A STATISTICAL ANALYSIS OF ENDOGENEITY AND INSTRUMENTAL VARIABLES OF EDUCATION ON INCOME IN NAMIBIA

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN APPLIED STATISTICS AND DEMOGRAPHY

OF

THE UNIVERSITY OF NAMIBIA

BY

MARIA KANDIWAPA NANYEMBA

200008765

APRIL 2024

MAIN SUPERVISOR: PROF. LAWRENCE KAZEMBE (UNIVERSITY OF NAMIBIA)

CO-SUPERVISOR: DR. OPEOLUWA OYEDELE (UNIVERSITY OF NAMIBIA)

ABSTRACT

Education is a well-known driver of income and a causal factor for economic prosperity and social development in modern society. It provides one with capital such as qualification, knowledge and skillsets relevant to secure stable gainful employment, as well as increase one's income, improve health conditions and general welfare. However, the measurement of the influence of education most often suffers from endogeneity suspicion due to the possible presence of a correlation between the education variable and the error term. This presents a critical issue when performing a cause-and-effect relationship analysis as neglecting endogeneity can lead to unreliable estimation results even when dealing with large samples because estimators of the model parameters will be inconsistent. To date, most studies that focuses on analysing the effect of education in Namibia have done so with other factors such as fertility and not income, while most of these studies used regression approaches such as the linear regression and logistic regression methods to perform their respective analyses. However, no study has simultaneously tried to account for the endogeneity of both education and income in Namibia. Thus, using the 2015/2016 Namibia Household Income and Expenditure Survey as well as accounting for the possible presence of endogeneity and for non-linear effects of continuous observed confounders, this study aimed at estimating the impact of education on income in Namibia, in addition to exploring different modelling strategies to capture the relationship between income and education, and other socio-economic and socio-demographic factors.

A total of 495,311 households containing 2,250,122 individuals aged 18 to 60 years who have been to school and have earned salary and/or wage as their source of income in the last 12 months of the survey period were considered in this study. Three different modelling strategies were carried out, namely: (i) Basic Ordinary Least Squares (OLS) regression without considering the clustering, stratification and weighting factors, (ii) Basic OLS regression considering the clustering, stratification and weighting factors, and (iii) Instrumental Variables & Two Stage Least Squares (IV-2SLS) regression, while the best fit modelling strategy to use for exploring the relationship between income and education, and other socio-economic and socio-demographic factors was determined using the Durbin and Wu-Hausman test of endogeneity.

Results from this study showed that majority (82.1%) of the households in Namibia derived their source of income from salary and/or wage in 2015/2016. The IV-2SLS was identified as the best modelling strategy to use since the Durbin and Wu-Hausman test of endogeneity confirmed the presence of endogeneity (i.e, education must be endogenous), thus, making the OLS strategies unreliable to use. Furthermore, factors such as education, age, sex, marital status, type of work, employment period, total hours worked and transport mode to work had positive and significant influences on income, while factors such as region, dwelling unit type, main material for the wall, household size, and no contribution to investment had a negative and significant influences. However, the type of employer had a mixed effect on income. In addition, the IVs estimators revealed that the average rate of return of schooling was 13% which was higher compared to the OLS estimates of 5%.

It is therefore recommended that the Namibian government and policy makers consider boosting more funding into education, especially in communities within the rural and underdeveloped regions still having traditional dwelling households whose structures are made of corrugated iron/zinc sticks, wooden poles, clay and/or cow-dung. In addition, as part of their poverty alleviation strategies, the government should ensure the distribution of equitable resources, particularly the educational resources, to enhance the development of rural and underdeveloped regions so that they too can benefit from high-quality education opportunities in the country.

Key words: Income, education, instrumental variables, endogeneity, ordinary least squares, two stage least squares, Namibia

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

2SLS	Two Stage Least Squares
2SGAM	Two Stage Generalized Additive Model
AIC	Akaike Information Criterion
DEFF	Design Effect
GAM	Generalized Additive Model
GLM	Generalized Linear Model
GMM	Generalized Method of Moments
ILO	International Labour Organization
IV	Instrumental Variables
IV-2SLS	Instrumental Variables & Two Stage Least Squares
NDPs	National Development Plans
NHIES	Namibia Income and Expenditure Survey
NSA	Namibia Statistics Agency
OLS	Ordinary Least Squares
SEM	Structural Equation Modelling
SRS	Simple Random Sampling

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God Almighty for granting me the wisdom and strength to undertake and complete this research work. Give thanks to the Lord, for He is good. His love endures forever (Ps 136:1).

I would like to express my gratitude to my primary supervisor, Prof. Lawrence Kazembe, and co-supervisor, Dr. Opeoluwa Oyedele, who guided me throughout this project. I gratefully acknowledge their valuable critiques, good judgement and supervisory work in trying to perfect this research report. I am extremely grateful for the patient guidance and courage to proofread countless pages of statistical formulas through. Completing this work would have been even more difficult were it not for my supervisors' support and assistance provided by other staff members of the Department of Computing, Mathematical & Statistical Sciences at the University of Namibia. I am indebted to them for their help.

To my lovely husband, Johannes Ndeyapo Nanyemba, thank you for your unwavering support and encouragement throughout my studies, and thanks for standing by me. To my georgeous daughters, Eufemia Nambunga Endjala and Hileni Soondaha Nangula; my sisters, Rachel Mambimbo and Meke Iyaloo Kapepo, thank you so much for the moral support, love and care you have provided me with during my study, I could not ask for more. I wish also to express my sincere gratitude to all my family**,** friends who supported me and offered moral support. Special thanks go to all my fellow classmates in Masters in Applied Statistics and Demography for soldiering on and encouraging one another during our study period.

Finally, I would like to thank the Namibia Statistics Agency for the financial support and for availing the data on the data catalogue online for all users to have free access.

DEDICATION

This research report is dedicated to my late mother, Albertina Nangolo Pinehas (June 2011) and late grandmother, Rakkel Nangula Nowa (October 2020), in their loving memories for their encouragement throughout my schooling years.

DECLARATIONS

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

No part of this thesis/dissertation may be reproduced, stored in any retrieval system, or transmitted in any form, or by means (e.g. electronic, mechanical, photocopying, recording or otherwise) without the prior permission of the author, or The University of Namibia in that behalf.

I, Maria Kandiwapa Nanyemba grant The University of Namibia the right to reproduce this thesis in whole or in part, in any manner or format, which The University of Namibia may deem fit.

Maria Kandiwapa Nanyemba ………………………………	………………………	02/04/2024 ………………
Name of Student	**Signature**	**Date**

CHAPTER 1: INTRODUCTION

1.1 Background of the study

Namibia is currently classified as an upper-middle income country with a per capita gross national income (GNI) between US$1005 and US$12,235, according to the World Bank's classifications (Anon, 2020), yet the country manifests high level of income inequality, with a Gini coefficient estimated at 0.57. The Namibian economy registered an average growth rate of 4.05% with the country's GDP per capita increasing from US$2,392 in 1980 to US$5,231 in 2017 (National Planning Commission, 2018). Although the Namibian economy experienced economic growth over the past two decades, this economic growth has not led to a significant reduction in income inequality. Income inequality in Namibia has shown a declining trend but remains very high compared to countries that have medium human development and Sub-Saharan Africa countries (Vo et al., 2019). Income inequality affects countries across all levels of development and with varying demographic, sociological and economic characteristics (Dabla-Norris et al., 2015). High and sustained levels of inequality, especially inequality of opportunity can entail large social costs. Entrenched inequality of outcomes can significantly undermine individuals' educational and occupational choices (Kumhof et al., 2015). Higher inequality lowers growth depriving the ability of lower-income households to stay healthy and accumulate physical and human capital. For instance, it can lead to under- investment in education as poor children end up in lower-quality schools and are less able to go on to colleges and universities. As a result, labour productivity could be lower than it would have been in a more equitable world (Kaasa, 2005).

A growing body of evidence suggests that rising influence of the rich and stagnant incomes of the poor and middle class have a causal effect on crises, and thus directly hurt short- and long-term growth (Dabla-Norris et al., 2015). In particular, studies have argued that a prolonged period of higher inequality in advanced economies was associated with the global financial crisis by intensifying leverage, overextension of credit, and a relaxation in mortgage-underwriting standards, and allowing lobbyists to push for financial deregulation (Zare, 2019). Higher top income shares coupled with financial liberalization, which itself could be a policy response to rising income inequality, are associated with substantially larger external deficits (Kumhof et al., 2015). Such large global imbalances can be challenging for macroeconomic and/or financial stability and growth (Bernanke, 2020; Royal et al., 2022).

Extreme inequality may damage trust and social cohesion and thus is also associated with conflicts, which discourage investment. Conflicts are particularly prevalent in the management of common resources where, for example, inequality makes resolving disputes more difficult (IMF, 2017). More broadly, inequality affects the economics of conflict, as it may intensify the grievances felt by certain groups or can reduce the opportunity costs of initiating and joining a violent conflict. In addition to affecting growth drivers, inequality could result in poor public policy choices, enhanced power by the elite could result in a more limited provision of public goods that boost productivity and growth, and which disproportionately benefit the poor (Doumbia & Kinda, 2019). Inequality hampers poverty reduction and affects the pace at which growth enables poverty reduction. Growth is less efficient in lowering poverty in countries with high initial levels of inequality or in which the distributional pattern of growth favours the non- poor (IMF, 2017). Moreover, to the extent that economies are periodically subject to shocks of various kinds that undermine growth, higher inequality makes a greater proportion of the population vulnerable to poverty.

Moreover, another important factor when it comes to income is education. Education can play an important role in reducing income inequality, as it determines occupational choice, access to jobs, and the level of pay, and plays a pivotal role as a signal of ability and productivity in the job market (Wolla & Sullivan, 2017). The impact of education as a determinant of income inequality (endogeneity) has been extensively studied elsewhere (Block et al., 2012) but not comprehensively in Namibia to explain the growing inequalities. The study by Jellenz et al., (2020) focused on the linkages between tertiary education parameters and Namibia's sustainable economic development. They concluded that the diversification of economic sectors that would result from raising the educational quality and improving access to Namibia's tertiary education system would hasten the country's process of internationalisation. Additionally, Namibia must contend with a host of issues that affect its tertiary education system, such as corruption, unemployment, and multifaceted poverty.

Education is expected to provide one with capital such as qualification, knowledge and skillsets relevant to secure stable gainful employment, increase one’s income, and improve health conditions and general welfare. Further, it allows allocation of resources efficiently to maximise output, contribute in terms of taxes to the nation and improve the quality of life of the individual (Marbuah, 2016). Ideally, education is one of the basic needs for human development and to escape from poverty (Sivakumar & Sarvalingam, 2015) it is necessary for national development and a prosperous society. According to Belfield and Levin (2007) a person’s educational attainment is one of the most important determinants of one's life chances in terms of employment, income, health status, housing, and many other amenities.

Endogeneity may arise due to the omission of explanatory variables in the regression, which would result in the error term being correlated with the explanatory variables, thereby violating a basic assumption behind ordinary least squares (OLS) regression analysis (Abdallah et al., 2015). It may also be caused by the dependent variable being influenced by one or several explanatory variables, which in turn are influenced by the dependent variable. Such endogeneity may be of the simultaneous type whereby contemporaneous realizations of both the dependent variable and the explanatory variables in question affect each other (Ketokivi & McIntosh, 2017). Consequently, it may be of the dynamic type whereby past realizations of the dependent variable influence current realizations of one or more of the explanatory variables. In most data sets, the variables of interest are often both cross-sectionally and serially correlated. As such, the common assumption of independence in regression errors is generally violated (Abdallah et al., 2015).

Broadly speaking, endogeneity concerns arise in situations where an explanatory or independent variable correlates with the error term (residual) of a specified model, rendering estimates inconsistent (Rutz & Watson, 2019). This is because the coefficient estimate of the compromised explanatory variable also contains the effect of unaccounted for variable(s) that also partially explain the dependent variable. If the possible presence of endogeneity is not accounted for, it results in major consequences and severity of the biases (Bascle, 2008). The large majority of studies use ordinary least squares estimation whose selection on observables assumptions are unlikely to hold (Kolstad & Wiig, 2015). The measurement of the influence of education, however, suffers from endogeneity suspicion because there may be a correlation between the education variable and the error term. Henceforth, neglecting this endogeneity can lead to unreliable estimation results even in large samples because estimators of the model parameters are inconsistent (Block et al., 2012). Thus, this study focused on modelling the impact of education on income using the 2015/2016 Namibia Household and Expenditure Survey accounting for the possible presence of endogeneity and for nonlinear effects of continuous observed confounders on income in Namibia.

1.2 Statement of the problem

Early human capital theory indicates that education may increase income, and thus, educational expansion or increasing public expenditure on education will decrease income inequality (Teweldemedhin, 2015). Thus, a positive relationship between income and education is expected, both at individual and country level, because more highly educated employees earn higher income, while collectively reflects similar pattern at macro-level. However, estimating effects of educational have often been contradictory or inconclusive, largely because the complex relationships between income and education have been neglected. It is likely that some factors, both observed and unobserved, that influence income may also influence the education/level of schooling. If that is the case, there will be correlations between these background characteristics and the error term, giving rise to the problem of endogeneity, thereby violating the independence assumption of ordinary least squares (OLS) estimation. Therefore, neglecting endogeneity can lead to unreliable estimates even in large samples because estimators of the model parameters are inconsistent (Block et al., 2012). Thus, it poses an immediate, fundamental threat to both the theoretical validity and the practical relevance of research findings (Ketokivi and McIntosh, 2017).

A flexible approach, two-stage generalized additive model (2SGAM) which involves applying a set of instrumental variables to the troublesome regressors before conducting the regression of theoretical interest is proposed to examine how strongly education influence income, through joint modelling of income and education to account for endogeneity, while adjusting for both observed and unobserved confounding non-linear effects (Zanin et al., 2015).

To date, most studies that focuses on analysing the effect of education in Namibia have done so with other factors such as fertility and not income, while most of these studies used regression approaches such as the linear regression and logistic regression methods to perform their respective analyses. However, no study has simultaneously tried to account for the endogeneity of both education and income in Namibia. Thus, using the 2015/2016 Namibia Household Income and Expenditure Survey (NHIES) as well as accounting for the possible presence of endogeneity and for non-linear effects of continuous observed confounders, this study aimed at estimating the impact of education on income in Namibia, in addition to exploring different modelling strategies to capture the relationship between income and education, and other socio-economic and socio-demographic factors. Findings from this study could provide useful input on the effect of education on individual income.

1.3 Objectives of the study

The main objective of this study was to model the impact of education on income using the 2015/2016 NHIES. The specific objectives were:

(a) To estimate the impact of education on income in Namibia accounting for the possible presence of endogeneity and for non-linear effects of continuous observed confounders.

(b) To explore different modelling strategies to capture the relationship between income and education and other socio-economic and socio-demographic factors.

(c) To examine the effects of important factors (covariates) that were simultaneously influencing income and education of the population in Namibia.

1.4 Significance of the study

The study of income inequality and education as a determinant is critical towards accelerated economic growth and development. Educational factors such as higher educational attainment and more equal distribution of education play a significant role in making income distribution more equal. The findings may assist in aligning education policies for increasing access for all, reduce income inequality, and escape the vicious circle of poverty for the Namibian populace. All in all, the findings from the study (models, conceptual frameworks/concepts and hypotheses) would add value to the body of scientific knowledge on education as an endogenous factor in Namibia and globally.

1.5 Limitation of the study

The findings of this study were however limited to the cross-sectional nature of the NHIES data used herein. Since this data was from a secondary source, the researcher did not have any direct influence on the questions asked and information elicited, thus any errors incurred during the data collection stage might have affected the outcome of this study. Additionally, this study was limited to only a single wave of survey which might not allow for comparison with earlier surveys and only to respondents who have earned salaries and wages as basic income during 2015/2016.

Moreover, this study only focused on basic income in the form of salaries and wages but excluded other forms of income such as subsistence farming, commercial farming, business activities, pensions from employment and/or annuity funds, cash remittances, rental income, interest from savings/investments, state old-age pension, war veterans/ex-combatants subvention, disability grants for adults (over 16 years), state child maintenance grants, state foster care grant, state special maintenance grants (disabled under 16 years), alimony and similar allowances, drought relief, and in kind receipts. Furthermore, the study's analysis was restricted to the 2015/2016 NHIES data collected over a period of 12 months at household- level for respondents who were aged 18 to 60 years in Namibia.

Income could be influenced by more than one endogenous variable other than education only, and the endogeneity could have been caused by other forms of endogeneity such as measurement error while this study only looked at the omitted variables which was dealt with via the Instrumental Variables (IV) and Two Stage Least Squares (2SLS) regression approaches. However, the data did not contain all the necessary variables, such as mother's/father's education and the background of household members, to see how they would have influenced education.

1.6 Organisation of the thesis

This thesis is structured into five chapters, summarized as follows: Chapter 1 covers the introduction, background of the study, significance of the study, problem statement and research objectives. Chapter 2 reviews and discusses existing relevant and empirical literature on problems of endogeneity. The research methodology used including research designs are outlined in Chapter 3. The results are presented and reported in Chapter 4. Chapter 5 discusses the key findings of the study and drawing of conclusion and recommendations.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter reviews past studies/methodology on endogeneity. Section 2.2 describes the income, education and other socio-economic factors, while Section 2.3 presents an overview of endogeneity. Section 2.4 outlines the depictions and description of the causes of endogeneity, with other approaches to address endogeneity given in Section 2.5.

2.2 Income, education and other socio-economic factors

Primarily, measuring endogeneity is not easy itself as various studies often used methods to address endogeneity (without explaining why endogeneity was a concern), and then proceed to justify how their used methods addressed the endogeneity concern, or providing adequate statistical evidence that supports the use of their approaches (Bliese et al., 2020; Li et al. 2021). Research indicates approximately 90% of papers published in premier journals have not adequately addressed endogeneity bias (Ullah et al., 2018). Confirmed by Antonakis et al. (2014), it was claimed that "researchers fail to address at least 66% and up to 90% of design and estimation conditions that make causal claims invalid" and this was based on a study conducted over 100 articles in top journals. Furthermore, the study by Hill et al. (2021) identified and reviewed 435 articles mentioning endogeneity. Of those, 86 (20.9%) addressed the omitted variable, 85 (20.7%) simultaneity, 6 (1.5%) measurement error, 132 (32.1%) selection and 102 (24.8%) were not clear. Despite that, the review of articles showed that, independent of methods being used incorrectly, many studies were unclear in discussing endogeneity.

2.2.1 Education and Income

Education is a factor for economic prosperity and social development in modern society. Education plays a vital role in many countries' economies, as well as the general health of labour markets and overall employment. Indeed, employment tends to go in lock step with increasing education standards. Marbuah (2016) expressed that one gains knowledge, skill sets, and the proper mindset through education, which can improve their lives. With these qualities, one can get employment, make money, maintain good health, get credit, become more productive, end poverty, and engage fully in society. Moreover, education allows its owner to receive a higher income and gives the opportunity for self-expression, creative fulfilment, as well as moral satisfaction from current activities (Stryzhak, 2020). The International Labour Organization Glossary of Statistical Terms defined education as an investment in human capital while income was defined as the revenue a business earns from selling its goods and services or the money an individual receives in compensation for his or her labour, services, or investments.

Endogeneity bias may provide erroneous estimates and inferences, which could result in misleading results and improper theoretical interpretations (Ullah et al., 2018). Using the Ordinary Least Squares (OLS) method, the results of the study conducted by Schultz, Tan, & Walsh (2010) showed no causal relationship between corporate governance and firm performance. Also, false correlations have shown that the prior substantial relationships discovered by the pooled OLS and fixed-effects models were still present (Jalil & Idrees, 2013). This further showed that: (i) a bias in the parameter estimates caused by dynamic endogeneity, unobservable heterogeneity, and/or simultaneity was the cause of the causal linkages discovered by the pooled OLS and the fixed-effects panels' models, and (ii) study conclusions can be sensitive when various econometric methodologies were used.

2.2.2 Education and health

In terms of determinants of socio-economic welfare, research has demonstrated that, at the individual level, measures of educational attainment and literacy skills were associated with health behaviours, health outcomes, and appropriate use of health services throughout one's life (Siddiqi et al., 2012). Likewise, healthy and educated people contribute more to economic growth, while health and education enable people to improve and use their capabilities. However, deprivation in health and education can affect people very severely. Education is one of the basic needs for human development and to escape from poverty, thus the deprivation of education in itself represents poverty (Sivakumar & Sarvalingam, 2015). Zanin, Radice, & Marra (2015) performed an endogeneity test on the impact of women's education on fertility using the 2010 Demographic and Health Survey in Malawi. The p-value obtained from the test provided evidence for them to reject the null hypothesis of no endogeneity and to conclude that a more complex modelling approach would be justified over a simpler model.

2.2.3 Education and wealth

The relationship between education and wealth is strong. Earning a higher income makes saving easier, and saving is necessary to build wealth (Wolla & Sullivan, 2017). Those with lower incomes have a flatter (non-humped) income pattern, which makes saving and paying down debt more difficult. However, those with more education also tend to make financial decisions that contribute to building wealth.

2.3 Overview of Endogeneity

Endogeneity of regression predictors arises as a result of issues like omitted confounding variables, simultaneity between a predictor and the outcome, and errors in regression covariates. This phenomenon is common in many areas of applied economics, including health economics and health services research studies which heavily rely on observational data(Ketokivi & McIntosh, 2017; Semadeni, 2020). In general, endogeneity issues appear when an independent or explanatory variable correlates with the residual error. A given model's $\mathbf{E}[\mathbf{x}|\mu] = \mathbf{cov}(\mathbf{x}'\mu) \neq \mathbf{0}$ can cause estimates to be inconsistent (Rutz & Watson, 2019), especially when the goal of such model is to calibrate the relationship between inputs (independent) variables and outputs (dependent) outputs. This is because the coefficient estimate of the compromised explanatory variable also contains the effect of unaccounted for variable(s) that also partially explain the dependent variable (Schultz et al., 2010). The estimated coefficients from an OLS estimation can be biased such that

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\varepsilon}$$

$$\mathbf{E}[\mathbf{b}] \neq \boldsymbol{\beta}.$$

A useful two-equation system which serves as a starting point for handling these issues can be given as follows:

$$y_i = f(s_i) + \beta x_i + \varepsilon_i$$

$$s_i = z_i\theta + u_i$$

$$\begin{bmatrix} \varepsilon_i \\ u_i \end{bmatrix} | x_i, z_i \overset{i.i.d}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \sigma_{\varepsilon u} & \sigma_u^2 \end{pmatrix}\right] \ , i = 1,2, \dots n$$

In the above system, s is a potentially endogenous variable, and the endogeneity (or unobserved confounding) problem is handled by introducing a possible correlation between u and ε , denoted by $\rho_{\varepsilon u} \equiv \sigma_{\varepsilon u}/\left(\sqrt{\sigma_u^2 \sigma_\varepsilon^2}\right)$. Figure 1 shows a general illustration of endogeneity.

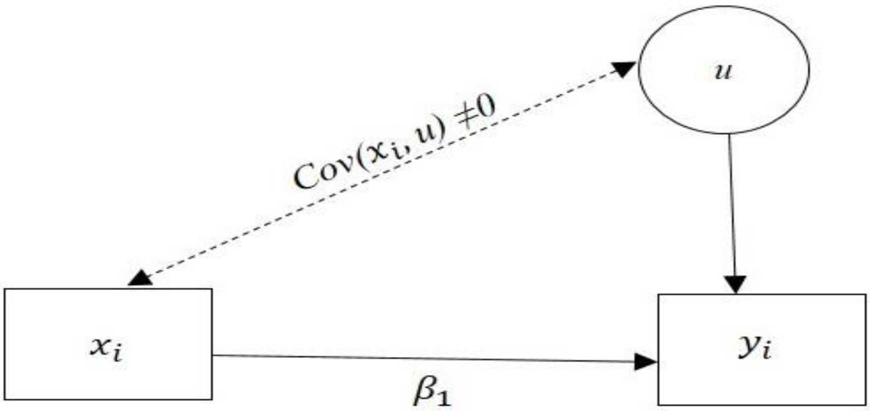

Figure 1: A schematic of a general endogeneity (Hill et al., 2021)

In general, endogeneity happens when the residual u arises because factors that influence y are not included in the regression function. When the predictor, x, correlates with the residual, u, this causes bias in the estimates of β . The bias can be upward or downward depending on the unmodelled factors that correlate with x and predict y as shown in Figure 1. Bliese et al. (2020) stipulates that the independent variable's contribution to the variance is not clearly defined since omitted factors that are related to both the independent and dependent variables may contribute to the relationship in some way. Finding the proportion of variance that is specific to the independent variable in the presence of endogeneity is difficult.

Wooldridge (2016) separates the causes of endogeneity into three categories: omitted variable, simultaneity and measurement error. Moreover, Wintoki et al. (2012) categorizes three potential sources of endogeneity, namely unobserved heterogeneity, simultaneity and dynamic endogeneity. Furthermore, Hill et al. (2021) presented mostly the three conditions of endogeneity detailing them with depictions and descriptions.

2.4 Depictions and Description of Three Causes of Endogeneity

Below follow three examples demonstrating the three most types of endogeneity using the example of measuring the impact of education on determining the income (salary).

(a) Omitted Variable

Due to uncertainty surrounding the exclusion of explanatory factors, one of the most frequent sources of endogeneity is also one of the hardest to diagnose (Rutz & Watson, 2019). This source of endogeneity arises from the omission of an explanatory variable that relates to the dependent variable as well as one or more other explanatory factors that were included in the analysis.

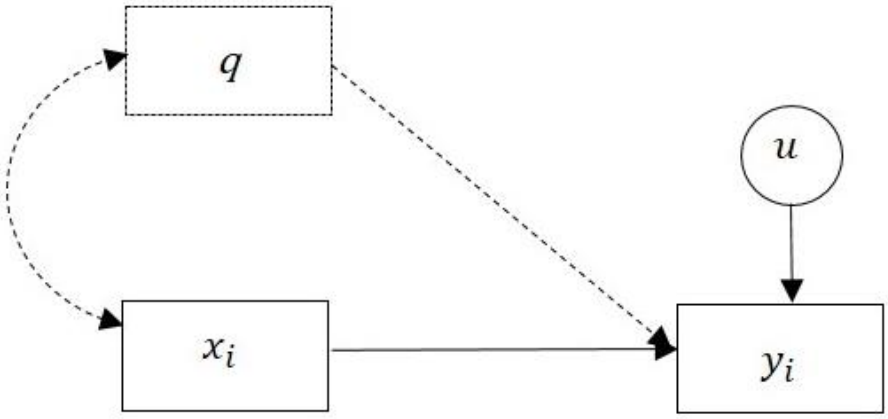

Figure 2: A schematic of an endogeneity with omitted variable (Hill et al., 2021)

Since the variable q is not modelled even though it affects y, residual u does not contain it. When q and x are also linked, q and the residual u are likewise associated, as illustrated in Figure 2. This research study focused on this kind of endogeneity (omitted variable). Hill et al. (2021) proposes multiple approaches to address omitted variable type of endogeneity. Endogeneity bias is essentially caused by multiple origins, and there are numerous ways to correct it (Wintoki et al., 2012). For instance, to handle panel data (i.e., dynamic endogeneity bias), the dynamic generalized method of moments model (GMM) is employed, and two-stage least squares (2SLS)/three stage least squares (3SLS) are frequently used for survey data (Li et al., 2021). The design/experiment, control and proxy variables, fixed effects, and instrumental factors are some of the approaches. Additionally, several methods, including the generalized method of moments (GMM), structural equation modelling (SEM), maximum likelihood, two-stage least squares (2SLS), and three-stage least squares (3SLS), can be used to estimate instrumental variables models. Hill et al. (2021), Rutz & Watson

(2019) and Semadeni et al. (2014) summarized the four (4) solutions to omitted variable endogeneity as follows:

i. *Design* - By randomly assigning participants to treatment and control conditions, experimental trials can avoid the endogeneity of omitted variables. Random assignment guarantees that, in an idealized format (i.e., with sufficient sample sizes, efficient manipulations, etc.), any omitted variable is evenly distributed across both conditions. Thus, the predictor will not display systematic variation with the residual.

ii. *Control and proxy variables* - This source of endogeneity is eliminated if the control variable measures the missing variable exactly. However, if the variable is unavailable and cannot be disregarded, another solution is to identify a substitute for the absent variable by changing an unobserved variable q with a substitute $\tilde{q}$, such that $y = a + Bx + \tilde{q} + u$.

iii. *Fixed effects* - The problem can be solved by estimating a model with individual or group fixed effects if an omitted variable is not available or directly observable, but theory or evidence indicates it is constant within a group or invariant over time. For each entity i in the analysis $y_i = a + c_i + Bx_i + u$, fixed effect terms will have the constant c_i added.

iv. *Instrumental variables* - A technique that relies on instrumental variables is frequently used to address omitted variable endogeneity. There might be a problem with missing data but measuring or using approximations to represent the variables is not practical. In these circumstances, instrumental variables provide a means of obtaining unbiased estimates, but their application necessitates the use of theoretical presumptions. In particular, instrumental variables call for the addition of a variable z that predicts the endogenous variable x in what is commonly referred to as "the first stage" $x = (a_x + B_x\text{z} + u_x)$ but is unrelated to the unexplained portion of the model u_y in what is commonly referred to as "the second stage" $\text{y} = (a_y + B_y\text{x} + u_y)$; subscripts indicate that coefficients and residuals differ in the two equations.

Since they must satisfy two requirements, finding suitable instrumental variables can be challenging. The first is relevance, which denotes that a variable is related to the endogenous predictor(s) x. You can directly test the presumption that the instrumental variable z has an impact on x. The second is exogeneity, which denotes that z is uncorrelated with the residual u of the result y, indicating that z's only impact on y comes from x. A researcher must instead present conceptual justifications for why the instrumental variable is uncorrelated with the residual in the second step of the regression because this second assumption—that z is exogenous—cannot be evaluated directly. The exogeneity criterion can be tested using so-called overidentifying restriction tests (like Sargan-Hansen or Sargan's -J), but only if the researcher has more instruments than are required. These tests are predicated on the model's correct identification. If the instrumental variable is legitimate, the Hausman test can establish whether an instrumental variable estimate method is required.

Supposed the true data generating model is given by:

$$y = \beta x + \gamma z + \varepsilon,$$

where,

y = income e.g. gross salary

x = job experience e.g. various skill set

z = one or more influential variables, e.g. years of schooling, cognitive ability

ε =i.i.d (independent and identically distributed) error term

Assume that the omitted variable is the cognitive ability. The β and γ above, are the parameters of interest. Yet, the researcher only observes y and x, omitting z, and then estimates

$$y = \beta x + \upsilon, \upsilon \text{ i.i.d error term}$$

Higher cognitive ability increases the likelihood that an individual would perform better at work and make more money. This holds true for both individuals with high and low levels of education. The average level of education for those with better cognitive abilities is higher. Incorrectly attributing all of the higher income to higher education levels when at least some of this should be attributed to higher cognitive ability occurs when indicators of cognitive ability are excluded from the regression. This leads to the problem where $\upsilon = z\gamma + \varepsilon$. If $\gamma \neq 0$ and x and z are correlated, then x is correlated with the error term υ, making x endogenous.

(b) Simultaneity

When one or more explanatory variables and the designated dependent variable in a model are causally related in a simultaneous and reciprocal manner, this is known as simultaneity in variables (Wooldridge, 2016).

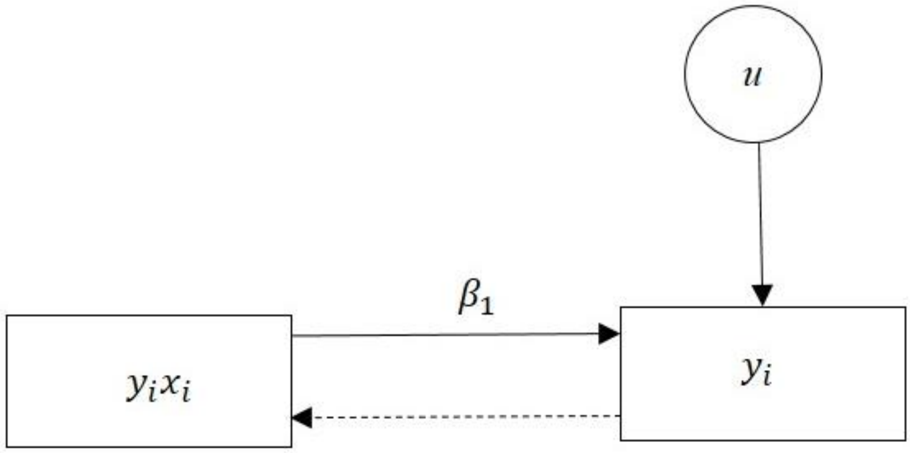

Figure 3: A schematic of an endogeneity with simultaneity (Hill et., 2021)

In addition to x causing y, Figure 3 demonstrates how y can also causes x, i.e., y will correlate with x, as will y and u. Only if u has zero autocorrelation can time separation address the issue (i.e., no correlation with itself over different time periods). The x at time 0 is not caused by u at Time 1 because it cannot cause something in the past; however, u at Time 1 is frequently correlated with u at Time 0 and as noted above, x at Time 0 is correlated with u at Time 0.

Suppose the true data-generating model, simultaneity takes the form:

$$y = \beta_1 x_1 + \gamma z + \varepsilon_1,$$

$$z = \delta_2 x_2 + y\emptyset + \varepsilon_2,$$

where

y = income e.g. gross salary

x_1 and x_2 = job experience at time $t = 1$ and $t = 2$

z = one or more influential variables, e.g. years of schooling, cognitive ability

$\varepsilon_1$and ε_2 = i.i.d error term (potentially correlated).

In this type of endogeneity, only y and x and z are observed, and then estimated as

$$y = \beta x + \gamma z + \upsilon, \upsilon \text{ i.i.d error term.}$$

The $E(z\upsilon) \neq 0$ is the consequence, making z endogenous. To make causal claims, one needs to disentangle the temporal order in which they influence each other.

(c) Measurement Error

The researcher finds difficulty in precisely estimating the true relationships between constructs when they are measured imperfectly or inconsistently if the measurement error in either dependent or the explanatory variables exist.

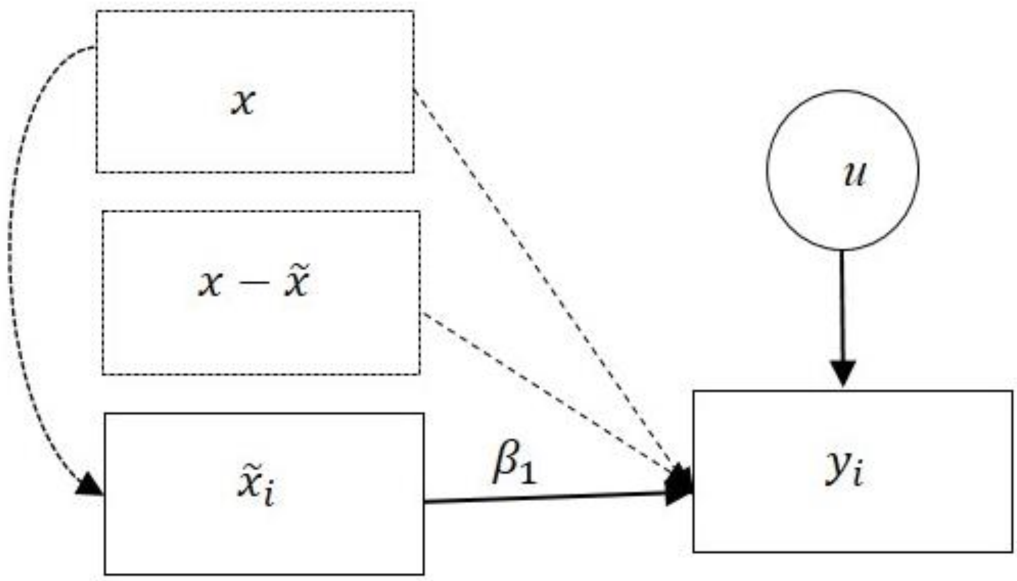

Figure 4: A schematic of an endogeneity with measurement error (Hill et al., 2021)

When there is measurement error, observe $\tilde{x}$ rather than x. As seen in Figure 4, when the measurement error in x influences y, it becomes a component of u, which is related to $\tilde{x}$ (Hill et al., 2021) such that:

$$y = \beta x + \varepsilon,$$

where,

y = income e.g. gross salary

x = job experience e.g. various skill set

ε =i.i.d error term.

Under this type, only income (y) without error and job experience (x), with measurement error of unknown origin are observed for job experience (x):

$$x = x + \eta, \eta \text{ is the measurement error,}$$

$$y = \beta x + \ \upsilon, \upsilon \text{ i.i.d error term}$$

As $\upsilon = \beta\eta + \ \varepsilon$ both x and υ are a function of η, making x endogenous.

The coefficient estimates, which depend on the variance of the explanatory variable in respect to the variance of the measurement error, will attenuate to zero when there is correlation between the observed explanatory variable and its measurement.

2.5 Other approaches to addressing endogeneity

It is important to note that there are various ways to mitigate endogeneity bias, which originate from various sources. For instance, to handle panel data (i.e., dynamic endogeneity bias), the dynamic generalized method of moments model (GMM) is employed, and two-stage least squares (2SLS)/three stage least squares (3SLS) are frequently used for survey data (Ullah et al., 2018). Two phases are involved in 2SLS estimation with instrumental variables. First, the selected instrumental factors are regressed on the endogenous independent variable, and the resulting regression residual is stored. The dependent variable is regressed on the residual rather than the endogenous independent variable in the second step (Zaefarian et al., 2017).

The approaches listed by Hill et al. (2021) to deal with endogeneity include (i) instrumental variable (ii) control function (iii) latent instrumental variable (iv) Gaussian copula (v) field experiment and (vi) structural model. Instrumental variables (IV) methods are a common family of strategies for dealing with such endogeneity. The portion of the endogenous variable that is not correlated with the error term is separated using an instrument, and this portion is then used to explain the dependent variable in the regression model. This research study focused on the instrumental variable with a two-stage least squares approach (2SLS). An instrument can only be related to the dependent variable through an endogenous variable. A strong (weak) instrument is one that has a high (low) correlation with the endogenous variable instrument (Rutz & Watson, 2019). The influence of the causal variable is exaggerated if the correlation between the omitted variable and the causal variable is positive; the effect is underestimated if the correlation is negative (Block et al., 2012).

The method of two stage least squares (2SLS) is second in popularity only to ordinary least squares for estimating linear equations in applied econometrics. The IV methods can be used to obtain consistent estimators in the presence of omitted variables. IV can also be used to solve the errors-in-variables problem, at least under certain assumptions (Hensher et al., 2018). Two conditions which are relevance and exogeneity need to be fulfilled for the instrumental variables. Relevance refers to the degree to which the instrument corresponds with the endogenous variable. A literature on instrument strength (i.e., strong vs. weak instruments) examines how relevance influences model results, and scholars have created recommendations based on the F-statistics of first-stage regressions to determine instrument relevance. The general conclusion of most research indicates that stronger (higher F-statistics) as opposed to weaker (lower F-statistics) instruments are better for two-stage approaches. Complementing instrument relevance, exogeneity refers to the degree to which an instrument is uncorrected with the disturbance term in the second stage.

2.5.1 Generalized Linear Models (GLMs) and General Additive Models (GAMs)

As mentioned earlier, GMM is used to address endogeneity when dealing with panel data while with instrumental variable (IV) addresses survey data. Highlights that the widely known linear regression model's principle is expanded by generalized linear models (GLM). Müller (2012) pointed out that the linear model assumes that the conditional expectation of the dependent variable, or response variable, is equal to a linear combination $X^T\beta$ i.e.

$$E(Y|X) = X^T\beta$$

This could be equivalently written as $Y = X^T\beta + \varepsilon$. Unfortunately, the linearity restriction does not allow for the consideration of many real-world circumstances. A continuous distribution of the error term, for instance, implies that the response Y must also have a continuous distribution. As a result, when working with binary Y or counts, the linear regression model may not succeed.

Udokang Anietie et al. (2015) mentions that a generalized linear model (GLM) consists of three components:

1. An unpredictable element that, given the values of the model's explanatory variables, specifies the conditional distribution of the response variable, Y_i (for the ith of n independently collected observations). The Y_i distribution in the original GLM formulation belonged to an exponential family, such as the Gaussian, binomial, Poisson, gamma, or inverse-Gaussian families of distributions.
2. A linear predictor - that is a linear function of regressors,

$$\eta_i = \beta_i X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

3. A smooth and invertible linearizing link function $g(.)$, which transforms the expectation of the response variable, $\mu_i = E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha = \beta_i X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

For $(y_i, x_i), i = 1, \ldots, n, let\ y_i$ be independently distributed in the exponential family and let $\mu_i = E(y_i)$.

Guisan et al. (2016) indicated that both GAMs and generalized linear models (GLMs) are built on the premise of an additive model that can simulate nonlinear dependence structures and the smoothness of the model fit such that

$$g(\mu_i) = f(x) = f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip})$$

for a link function $g(.)$. Using the link function $g(.)$, the GLM models the response variable's mean. The components of additive models are functions f_1 to f_p, each of which is a function of the predictor x. Additionally, a GLM presupposes that all the component functions will be linear. They expand GLM by letting the component functions be any arbitrary smooth functions, commonly approximated by using splines, which gives you more modeling options when there may be unidentified nonlinear dependencies in the response's mean (Gretz & Malshe, 2019).

Furthermore, a study by Zanin et al. (2015) highlights that it is crucial to take into consideration the possibility of both unobserved confounding effects and nonlinearities in the relationship between observed confounding effects and education when analyzing the connection between women's education and fertility. To address these problems, the study used a flexible model, the two-stage generalized additive model (2SGAM). This method eliminates the need to specify a priori the analytical form of the treatment-response and confounder response relationships (Zanin et al., 2015). Although GLMs and GAMs can be employed with assessment criterion like the Akaike Information Criterion, variable selection is essentially the same for all of the above regression models. In all models, predetermined guidelines such deviance reduction as indicated by the χ^2 statistic or methods that minimize AIC can be used (Antonakis et al., 2014).

Additionally, when a regression model is used to explain the dependent variable, an instrument separates the portion of the endogenous variable that is not linked with the error term (Reeb et al., 2012). The relevance criterion stipulates that an instrument must be correlated with an endogenous variable rather than the regression error term, i.e., it must be exogenous (Stock & Watson, 2019). Therefore, studies that treat education as exogenous by assuming the absence of unobserved confounding effects cannot determine whether education causally affects income. Estimates obtained will account for observed and unobserved confounding variables to better understand the relationship (Rutz & Watson, 2019; Schultz et al., 2010; Ketokivi & McIntosh, 2017).

Moreover, Zanin et al. (2015) modelled the impact of women's education on fertility in Malawi by employing a classic GAM approach which can model flexibility covariate effects but cannot account for unobserved confounders. By enabling the assessment of potential nonlinear effects of predictors on the response variable, generalised additive models (GAMs) go beyond generalised linear models (GLMs). The model structure of a GAM is presented by:

$$y = g^{-1}(\eta) + \epsilon, y \sim F \quad (1)$$

where

$$g^{-1}(\eta) = \mu = \mathbb{E}\,(y|X)$$

with $g(.)$ being a suitable link function; additive, ϵ is an additive, unobservable error trivially defined as $\epsilon \equiv y - g^{-1}(\eta)$ with $\mathbb{E}(\epsilon|X) = 0$; and F is an exponential family distribution. X can be written as (X^*, X^+) where $(X^* = (X_e^*, X_o^*, X_u^*)$ and $(X^+ = (X_e^+, X_o^+, X_u^+)$. The linear predictor of a GAM is typically given as follows:

$$\eta = X^*\beta^* + \sum_j f_j\,(x_j^+),$$

where $\beta^* = (\beta_e^+, \beta_o^+, \beta_u^+)$ represents the vector of unknown regression coefficients for $X^*$and the f_j are unknown smooth functions of the covariates x_j^+ in X^+, which can be approximated using, for instance, regression splines.

In the GAM context, the endogenous variables can be modelled by using the following set of flexible auxiliary regressions (Terza, 2016).

$$x_{ep} = g_p^{-1}\left\{Z_p^*\alpha_p^* + \textstyle\sum_j f_j(z_{jp}^+)\right\} + \xi_{up},\; p = 1, \ldots, h, \quad (2)$$

where x_{ep} represents either the p^{th} discrete or continuous endogeneous predictor, $Z_p^* = (X_o^*, X_{IVp}^+)$, with corresponding vector of unknown parameters α_p^*, and $Z_p^+ = (X_o^+, X_{IVp}^+)$. The reason why the regressions in Equation 2 can be used to correct the parameter estimates of the equation of interest is as follows. Once the observed confounders have been accounted for and if the IVs meet the afore-mentioned conditions, the ξ_{up} contain information about the unobserved confounders; for a more rigorous treatment, using a Taylor approximation of order 1, show that the components entering the ξ_{up} include unobserved variables. In fact, consistent estimates for the parameters in equation (1) can be obtained by using the following 2SGAM approach as follows:

(i). For each endogenous variable in the model, obtain consistent estimates of α_p^* and the f_j by fitting the corresponding equations (2) through a GAM method. Then, calculate the quantities

$$\xi_{up} = x_{ep} - \mathrm{g}_p^{-1}\left\{Z_p^* \alpha_p^* + \sum_j f_j\left(z_{jp}^+\right)\right\} + \xi_{up}, \qquad p = 1, \dots, h;$$

(ii). Fit a GAM defined by

$$y = \mathrm{g}^{-1}\left\{X_e^* \beta_{eo}^* + \sum_j f_j\left(x_{jeo}^+\right) + \sum_p f_p\left(\xi_{up}\right)\right\} + \zeta, \qquad (3)$$

where $X_{eo}^* = (X_{eo}^*, X_o^*)$, with parameter vector β_{eo}^* and $X_{eo}^+ = (X_e^+, X_o^+)$. The $f_p(\xi_{up})$ in equation (3) allows flexible account for the impacts of unobserved confounders on the response, the linear/non-linear effects of the endogenous regressors can be estimated consistently. Note that the $\hat{f}_p \hat{\xi}_{up}$ will contain a mixture of effects which makes them not interpretable, but this does not pose any problems since all that is required is to account for the presence of unobservables.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This chapter gives the methods that were used in the study to collect and analyse the data. Specifically, this chapter presents the research design, the population, the sampling techniques, the research instrument, data collecting procedures and the methods of data analysis. Finally, this chapter presents the ethical issues, which were considered in this study.

3.2 Research Design

This study used the secondary data from a household-based survey, 2015/2016 Namibia Household and Income Expenditure Survey (NHIES). The NHIES aims to give information to measure various socioeconomic indicators, such as actual patterns of consumption and income, as well as the degree of living conditions of the Namibian people. Additionally, the survey was created to give decision-makers access to quality, up-to-date, and trustworthy national, regional, and disaggregated statistics for rural and urban areas for planning and analysis. The first NHIES was conducted in 1993/1994, followed by the NHIES 2003/2004, then NHIES 2009/2010 and the latest NHIES 2013/2014. In 2003/2004 it was decided to align the undertaking of NHIES with the planning process of NDPs and conduct the survey at five-year regular intervals.

The sample design for the survey was a stratified two-stage cluster sample, where the first stage units were geographical areas designated as the Primary Sampling Units (PSUs) and the second stage units were the households. The up-to-date list of households in the selected PSU were prepared during the listing stage of fieldwork, and 12 households were systematically selected in each PSUs.

3.3 Population

The population of study was all individuals in the private households in Namibia aged between 18 and 60 years who earned a salary and/or wages as a main source of income in the year 2015/2016.

3.4 Sample

The sample from the 2015/2016 NHIES dataset for this study was the total of 495,311 households containing 2,250,122 individuals aged 18 to 60 years who have been to school and have earned salary and/or wage as their source of income in the last 12 months of the survey period were considered in this study.

3.5 Research Instruments

The original main questionnaire used in the data collection process of 2015/2016 NHIES consisted of 21 sections, but this study focuses on sections including education (Section 3) and income (Section 10). The improved methodology Computer Assisted Personal Interview (CAPI) that was used ensures efficient geo-coding of the questionnaires during data capturing and processing. This paperless method made it possible to in-build quality checks, edit rules and validation mechanisms into the application to control for data errors and inconsistencies interactively during the interview process. Such approach enhances timeliness, data integrity and reliability.

The 2015/2016 NHIES Data Entry application was built with many consistency checks, skipping patterns and other validations such as maximum and minimum acceptance range per variable. Supervisors were given minimum variables to check on a day-to-day basis. As a result, data consistency checks, coding and validation was done at field level. This minimized the time spent on post data cleaning, validation and editing process. The section on education is where the level of education, highest grade completed, years of study and literacy status of household members from age 6 and above are captured. The section on income sources and debts collected information about main and other sources of individual and household income. It also captured information about household debts (Namibia Statistics Agency [NSA], 2017).

3.6 Data Processing Procedure

The 2015/2016 NHIES was obtained from the NSA's Data Catalogue online. The focus was on individuals' income in households in the working age category with a special interest to those who have ever been to school and have earned salary and wage as a source of income, so the dataset was filtered to remain only with such cases. The completeness of the data was checked and it was found that some variables had incomplete data. This implies that missing cases were left out of the data analysis. Income variables are usually skewed variables with outrageous outliers, which would mean they are not normally distributed. Transformation was done such as squaring the income variable or taking a log of it to smoothen the variable out. Logarithmic transformation is a convenient mean of transforming a highly skewed variable into a more normalized dataset (Afifi et al., 2007). In this study, the income variable, being a dependent variable was transformed using the log to smoothen it.

3.7 Data analysis

Complex sample surveys involve the identification and data collection of a sample of population units via multiple stages or phases of identification and selection by means other than simple random samples. Stratified multistage cluster sampling introduces non-independence across units, disproportionate sampling, where some groups may be oversampled, and probability proportional to size sampling are examples of complex sampling designs. The complex design features (stratification, multi-stage clustering and unequal sampling weights) were considered. Failure to account for the stratification in the analysis will result in an overestimation of the standard error, and hence too wide confidence interval (Lu & Yang, 2012). The 2015/2016 NHIES sample design is of complex nature which means it includes (stratification, clustering (PSUs) and weights), hence the analysis took this into account during the survey set in the STATA software.

Furthermore, an approach for addressing disproportionate sampling and non-independence is to adjust the normalized weight by the design effect. A normalized weight is computed as the raw weight divided by its mean. The normalized weight takes the unequal selection probability into account but does so assuming a simple random sample. In other words, the non-independence has still not been addressed through a normalized weight, but this can be done by the inclusion of a Design Effect (DEFF). The DEFF provides a measure of degree of departure from SRS on the precision of a sample estimated. The DEFF is the ratio of the estimated variance derived from considering the sampling design to that derived from a simple random sample. A DEFF that is larger than 1.0 suggests that there is decreased precision of the estimate relative to what would have been obtained from a simple random sample, and thus DEFF values less than 1.0 indicate increased precision. Adjusting the normalised weight by the DEFF will produce more accurate standard errors than when the complex sampling features are simply ignored).

Moreover, in clustering, the total population was divided into some groups (or clusters) and a sample of the groups is selected. In multistage clustering, the clusters selected at the first stage are called primary sampling units or PSUs. Further sample selection occurs within PSUs and so on. In general, failure to account for the clustering in the analysis may lead to an underestimation of variability.

In this study, all the regression analyses performed were of multivariate nature and performed using STATA Version 16 software. Different sources of variability in the data collected at the individual level were explored. Variations in individual income between urban and rural were studied. This study further investigated the endogeneity and instrumental variables of education on income levels using the 2015/2016 NHIES.

3.7.1 Income & earnings

The income as the variable of interest in this study referred to the basic wage and salaries for an individual in the past 12 months. For the purpose of this study, this excluded allowances, commissions, bonuses and overtime payments. Earnings (i.e., wages and salaries), on the other hand, is the concept of earnings as applied in wages statistics, relates to remuneration in cash and in-kind paid employees for time worked or work done together with remuneration for time not worked, such as annual vacation and other paid leave or holidays (ILO, 2014).

The income recorded within the NHIES data was the gross income received before deductions. Income variables are usually skewed variables with outrageous outliers, making their distributions to be skewed with a long tail towards larger values. To produce a more symmetric distribution or smoothen the variable, transformation is usually applied on income variables such as squaring the income variable or taking a log of it etc., which makes the distribution closer to being symmetric, and yielding a better approximation to a normal distribution (Afifi et al., 2007). For this study, information about each respondent's income was determined through a specific question of Section 10: Income by individual which reads "Did [NAME] have any salary or wage income in the past 12 months?" **If yes, what was the salary and wage income for the past 12 months for [NAME] for the following:**

1. Basic monthly wage or salary

2. Bonuses and income from overtime

3. Sitting allowance (commissioners' and directors' fees)

4. Transport allowance

5. Housing allowance

6. Clothing allowance

7. Other cash allowance (specify). Please report income before deductions (gross income).

3.7.2 Education

Education was measured in terms of total number of years of schooling one has studied. This was asked through a question **"what is the total number of years [NAME] has studied (including repetitions)?".** The years of schooling is widely used as a proxy variable for education because knowledge and skills should increase with advancement in schooling. However, one limitation of this variable is that it does not measure performance in school, curriculum content, and other qualitative aspects of education (Zanin et al., 2015).The Namibia's current education system comprises 7 years of primary school (Grades 1 to 7), 6 years of secondary school (Grades 8 to 13) and 4 years of university education. Some factors that may affect education were not part of this analysis such as quality of the schools and teaching materials, the background of household members, preferences, abilities and motivation of the household members.

3.7.3 Observed confounders

The 2015/2016 NHIES data included a number of observed confounders that might affect household members' income and the relationship between education and income. The following variables were considered in the study:

Age of respondent – Age in completed years for the household member. The age variable was a continuous variable ranged between 18 and 60 years. This age range is typically considered to span for the working age population. In Namibia, a child is supposed to be in school until Grade 12 at around 17/18 years except in cases of child labour. An 18-year-old person is considered an adult and can make their own decisions to be in school or to look for a job to earn an income. For one to earn a salary/wage, one must have worked for a certain period to qualify to be paid a salary.

Contribution to investment - Ability of the household to contribute to investment.

Distance to the nearest high school - The distance that learners travel from home to the nearest high school measured in kilometers.

Dwelling unit type - Housing characteristic is an important indicator of household socio-economic status. The housing characteristic considered in this study was described with regard to the type of dwellings occupied by the household, including building materials used for the roof, floor and wall. This includes detached house, traditional dwelling, improvised housing unit and town house/apartment.

Employment period – The employee's period of employment with an employer at a particular time is the period of continuous service the employee has completed with the employer at that time as an employee. This was measured in numbers of years the employee has worked.

Highest level of educational attainment - Educational attainment refers to the highest level of education that an individual has completed.

Household member & size – To qualify as a household member, a person must have stayed in the household and not absent for a period six months or more in the past 12 months. Any person who was part of the household during the time of the survey and had met the household member requirement, this constitutes the household size.

Internet access –Whether the household has access to internet or not.

Main material for the wall – This variable looked at the main material for the outer walls of the main dwelling which comprises of three categories corrugated iron/zinc; sticks, wooden poles, clay and/or cow dung; others.

Marital status – The current civil status of each individual in relation to the marriage laws or customs of Namibia at the time of the interview. These categories are never married, married certificate, married traditionally, customary, consensual union, widowed, divorce and separated.

Region of residence (region) – A categorical predictor that identifies the geographical region of Namibia where each respondent lived during the time of interview. There are currently 14 regions in Namibia, namely; //Kharas, Erongo, Hardap, Kavango East, Kavango West, Khomas, Kunene, Ohangwena, Omaheke, Omusati, Oshana, Oshikoto, Otjozondjupa and Zambezi (see Appendix III).

Sex of respondent – The biological difference or the state of being male or female as differentiated with reference to the reproductive functions. The sex variable was coded as 1 for males and 2 for females in this study.

Source of energy for lighting – This refers to the main source of energy the household uses for lighting especially when household members are studying. The categories included electricity from main, battery lamp/torch/cell phone, candles, solar energy and others.

Transport mode – The type of transport mode household members used to the nearest facility e.g., primary school, high school, hospital or clinic, combined school, vocational training centre, university etc. For this study, only the high school facility was considered as those from 18 years and above would have deemed to have gone to high school.

Total hours worked – The number of hours all household members between 18 and 60 years have worked per day on the job in the past week.

Type of employer – The four types of employers all the salaried/wage employees worked for including private company, state-owned enterprises, commercial; government, public sector or army; private household/individual.

Type of work – The type of work (wage activities) all households members aged 18 years and above were classified to have worked, whether in agriculture or not in agriculture industry.

Urban/Rural area - Urban areas were defined as all proclaimed municipalities and towns in Namibia and the rest of the areas form part of rural area.

Table 1 shows the categories of variables considered in this study, including a brief description and their corresponding variable item within the NHIES dataset.

Table 1: Variables and Definitions

Variable	**Description/Definition**	**Variable in the dataset**
Quantitative variables		
Income	Salary/wage (transformed) to become log(salary/wage)	q10_02_1_log
Education	Measured in total number of years of study	q03_05
Distance to nearest high school	Distance in km to high school	q02_51_05
Household size	Number of household members	hh_members
Age of respondent	Age in single years	q01_06_y
Qualitative variables		
Region of residence (region)	14 regions of the country	region
Urban/Rural areas	Geographical type where household members live	urbrur
Dwelling unit type	Dwelling unit type	q02_01
Main material for the wall	Main material used for the wall	q02_03
Source of energy for lighting	Main source of energy for lighting	q02_09
Internet access	Household internet access	internet_access
Sex of respondent	Either respondent is male or female	q01_02
Marital status	Civil status of each individual in relation to the marriage laws or customs of the country	q01_07
Contribution to investment	Ability of the household to contribute to investment	q20_04

Type of employer	Type of employer household members work for	q06_29_1
Type of work	Type of work household members work in	q06_19_1
Total hours worked	Number of total hours worked in a day for the entire week	totalhrs
Transport mode	Type of transport mode household members used to go to the nearest facility	transport_mode
Employment period	Period how long household members are employed	employ_period

3.8 Modelling the Income-Education relationship - Instrumental Variable & Two Stage Least Squares (IV-2SLS) regression

To estimate the effect of education on income, the following model expression was used.

$$log(income) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j + \beta education + u_1$$

where income was the dependent variable, education was the variable of interest, x_j were the exogenous variables, β_0 is a constant, and u_1 is an error term with $E(u_1) = 0$. The above expression can be expanded into a multiple linear income model such that:

$$log(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \beta_E EDU + u_1 \qquad (4)$$

where $x_1, \ldots, x_j$ are exogenous variables (including past experience) and u_1 is a zero mean disturbance term. From the foregoing, $cov(x_j, u_1) = 0$ for $j = 1,2, \ldots, n$ but allow the possibility that EDU might be correlated with u_1. In other words, the explanatory variables $x_1, \ldots, x_j$ are exogenous, but EDU is potentially endogenous.

The IV estimator is known to be an appropriate estimator in the presence of these problems (Wooldridge, 2016). The idea of the IV approach is to estimate both equations simultaneously.

The IV approach exploits the existence of an *identifying* instrument, possibly a vector z_1, that is not captured in equation (4) that satisfies two conditions: (a) $cov(z_1, u_1) = 0$ and (b) $\theta_1 \neq 0$ in the reduced-form equation for the endogenous explanatory variable *EDU*, such that

$$EDU = \eta_0 + \eta_1 x_1 + \cdots + \eta_j x_j + \theta_1 z_1 + u_2 \qquad (5)$$

where $E(u_2) = 0$ and u_2 is uncorrelated with the $x_j (j = 1, \ldots, \ldots, j)$ for $j = 1,2, \ldots, j$ and z_1. Condition (a) relates to the *validity* of the (identifying) instrument(s), and condition (b) relates to the *quality* of the instruments. Basically, there need not to be a correlation between income and the instruments ($\theta_1 \neq 0$) *but then there is a correlation between the endogenous variable and the dependent variable* $cov(z_1, u_1) = 0$ (Wooldridge, 2016). For this study, the results from estimating equation (5), that is, the determinants of education are presented in Table 18 using both OLS and IV estimators. The first column of coefficients shows the (biased) estimation results that ensue when estimating equation (5) by OLS. This comparison underlines the practical importance of correcting for endogeneity biases when attempting to explain income.

Equations (4) and (5) were represented by the STATA code of**:** svy linearized : ivregress 2sls q10_02_1_log i.region i.dwelling_unit_type i.main_material_wall i.q20_04 q01_06_y i.q01_02 hh_members i.marital_status i.employer_type i.work_type i.totalhrs i.employ_period i.transport_mode (q03_05 = i.internet_access i.attain i.main_source_energy_lighting q02_51_05) (see Appendix 1 for more detailed codes). This code explains the instrumental regression from survey data with the complex design (svy linearized: ivregress 2sls) for the two-stage least squares (2sls) regression with the list of included exogenous/independent variables for the dependent variable (log income). In the brackets is the endogenous regressor (education) and the excluded exogenous variables.

3.9 RESEARCH ETHICS

This research study was carried out in accordance with the University of Namibia's Research Ethics Policy and Guidelines. The 2015/2016 NHIES data used in this study is freely available under the Central Data Catalogue at the NSA's website. Also, the data were already anonymized as per the NSA Microdata Statistics Act No. 9 of 2011. Additionally, since this data is freely & readily available on a public domain and downloadable from the NSA website, no separate permission is required for the data usage and publication.

CHAPTER 4: RESULTS

4.1 Introduction

Chapter 4 gives the results of the descriptive statistics, hypotheses tests, models fitted and the model selection criterion used.

4.2 Descriptive Statistics

Most households (82.1%) in 2015/2016 in Namibia derived their source of livelihood from salary and income, while only 0.2% relied on commercial farming as their survival source as shown in Table 2. From the Inter-censal Demographic Survey of NSA (2017), it was revealed that among households in the urban areas of the countries, 72% indicated salaries and wages as their main source of income, followed by business income with 11.3%, compared to households in the rural areas, where 31.8% of the households reported salaries and wages as their main source of income, followed by subsistence farming with 22.4%.

Table 2: Number of households by main source of income

Main source of income	Frequency	Percentage
Salaries & wages	406,593	82.1
Pension	15,438	3.1
Subsistence farming	15,307	3.1
Business income	32,397	6.5
Remittances/grants	15,277	3.1
Drought/in-kind receipts	3,746	0.8
Commercial farming	723	0.2
Others	5,830	1.2
Total	**495,311**	**100**

Source: 2015/2016 NHIES Household data

Moreover, looking at the mean salary for all 14 regions (shown in Table 3), it can be observed that households in the Khomas region had an average income of N$81,736.28 in 2015/2016, while households in the Kavango West region recorded the lowest average income of N$16,490.46.

Table 3: Mean income by region

Survey: Mean estimation				
Number of strata = 28	Number of observations = 7,843			
Number of PSUs = 841	Population size = 2,206,802			
	Design df = 813			
Region	Mean income	Linearized Std. Err.	[95% Conf. Interval]	
!Karas	63686.71	10078.56	43903.6	83469.8
Erongo	53842.76	5647.648	42757.1	64928.5
Hardap	47427.13	7848.522	32021.4	62832.9
Kavango East	26590.69	4787.212	17193.9	35987.5
Kavango West	16490.46	3940.24	8756.21	24224.7
Khomas	81736.28	8650.583	64756.2	98716.4
Kunene	47794.95	10112.67	27944.9	67645
Ohangwena	28017.98	6716.127	14835	41201
Omaheke	37512.55	6791.444	24181.7	50843.4
Omusati	32553.35	5530.756	21697.1	43409.6
Oshana	57185.92	10391.84	36787.9	77583.9
Oshikoto	54255.56	8331.365	37902	70609.1
Otjozondjupa	46029.85	5400.91	35428.5	56631.2
Zambezi	17623.87	4258.291	9265.33	25982.4

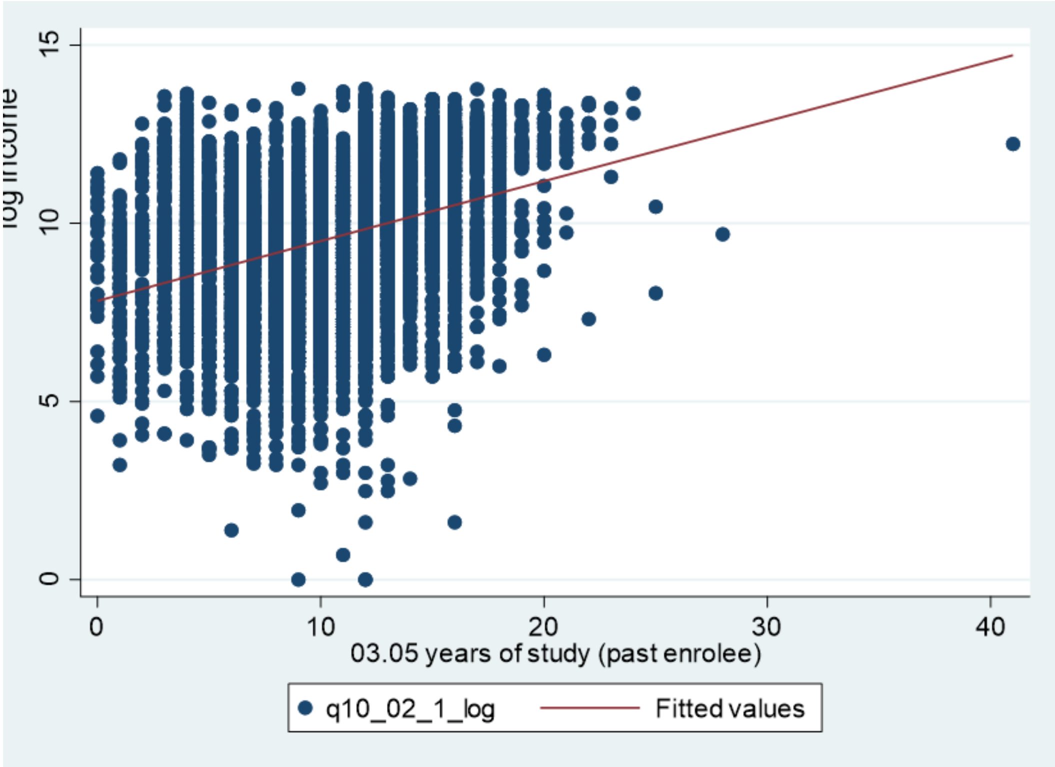

Figure 5: Scatter plot of Income against Education

Figure 5 shows the scatter plot of income against education. Most points fell along the fitted line, indicating a correlation between income and education exists and the dependent variable (income) may have multiple values for each value of the independent variable (education).

Table 4: Mean income by educational attainment in Namibia

Survey: Mean estimation				
Number of strata = 28		Number of observations =7628		Design df = 813
Number of PSUs = 841		Population size = 2137923		
		Linearized		
Educational attainment	**Mean**	**Std. Err.**	**[95% Confidence Interval]**	
No formal education	10696.51	1301.514	8141.791	13251.24

Primary	21274.46	1298.975	18724.72	23824.2
Secondary	45503.81	2497.996	40600.52	50407.09
Tertiary	151846.8	8543.581	135076.7	168616.8
Not stated	30727.79	16730.34	-2111.958	63567.54

Table 4 shows the mean income by educational attainment. Those with tertiary education had higher mean income than those with secondary and primary education. However, those with no formal education had the lowest mean income.

Table 5: Mean Income by sex in Namibia

Survey: Mean estimation				
Number of strata = 28	Number of observations = 7843			
Number of PSUs = 841	Population size = 2206802			
	Design df = 813			
Sex	Mean	Linearized Std. Err.	[95% Confidence Interval]	
Female	48515.13	2982.003	42661.8	54368.47
Male	62584.91	3767.69	55189.37	69980.46

From Table 5, it can be observed that on average, the male population of Namibia received a higher income than their female counterparts.

Table 6: Mean income by work type in Namibia

Survey: Mean estimation				
Number of strata = 28	Number of observations=7,288			
Number of PSUs = 832	Population size=2,022,039			
	Design df=804			
		Linearized		
	Mean	Std. Err.	[95% Confidence Interval]	
Work type				
Wage/ Salary Activity: In agriculture	25519.17	3482.338	18683.62	32354.72

Wage/ Salary Activity: Not in agriculture	63866.85	3672.885	56657.27	71076.43
Others	39168.4	9981.506	19575.51	58761.29

Those who earned wage/salary in industries other than agriculture (not in the agriculture) industry earned a salary almost double that of those who worked in the agriculture industry as indicated in Table 6 above.

Table 7: Mean income by employment period in Namibia

Survey: Mean estimation				
Number of strata = 28	Number of observations = 7,041			
Number of PSUs = 829	Population size = 1,950,575			
	Design df = 801			
		Linearized		
	Mean	Std. Err.	[95% Conf.	Interval]
Employment period				
Less than 1 year	16779.56	1866.452	13115.85	20443.28
1-5 years	54579.52	3289.75	48121.97	61037.07
6-10 years	101935	8411.911	85423.03	118447
11 years or more	119796.9	7134.463	105792.5	133801.4

Table 7 indicates various mean income groups as per their employment period. Those who had been employed for more than 11 years earn seven times higher than those who had been in employment for less than a year and earned almost double than those in employment between 1 to 5 years.

Table 8: Mean income by type of dwelling unit in Namibia

Survey: Mean estimation				
Number of strata = 28		Number of observations =7842		
Number of PSUs = 841		Population size=2206777		
		Design df=813		
		Linearized		
Type of dwelling unit	Mean	Std. Err.	[95% Conf. Interval	
Detached house	78136.7	5021.906	68279.27	87994.13
Traditional dwelling	21142.03	1943.021	17328.1	24955.96
Improvised housing unit	22760.25	1804.775	19217.68	26302.82
Town house/Apartment	97278.27	8653.043	80293.33	114263.2
Others	28738.04	4575.697	19756.47	37719.61

Table 8 shows the mean income of household members by type of dwelling unit they live in. Those who owned town house/apartment earned higher income than those in traditional dwelling and improvised housing units.

Table 9: Mean income by main material for the wall in Namibia

Survey: Mean estimation				
Number of strata = 28	Number of observations = 7842			
Number of PSUs = 841	Population size =2206777			
	Design df = 813			
		Linearized		
	Mean	Std. Err.	[95% Conf.	Interval]
Main material for the wall				
Cement blocks/Bricks/Stones	82364.83	4639.132	73258.75	91470.92
Corrugated iron/Zinc	22859.16	1725.16	19472.87	26245.46
Sticks, wooden poles, clay and/or cow-dung	12936.68	1504.504	9983.506	15889.85
Others	56517.75	9655.233	37565.63	75469.88

It can be observed from Table 9 that mean income for households whose dwelling units' main material for the wall was cement blocks/bricks/stones was six times higher than the mean income for households whose dwelling units' main material for the wall was sticks, wooden poles, clay and/or cow-dung.

Table 10: Mean income by total hours worked in Namibia

Survey: Mean estimation				
Number of strata = 28	Number of observations = 5,976			
Number of PSUs = 810	Population size = 1,628,297			
	Design df = 782			
	Mean	Linearized Std. Err.	[95% Confidence Interval]	
Total hours worked				
1-8	17953.13	6365.715	5457.219	30449.04
8+	64434.92	3748.478	57076.65	71793.19

Table 10 shows the mean income by total hours worked. The mean income of the population that worked more than 8 hours, was triple the income of those who worked between 1 and 8 hours a day.

4.2.1 Model 1 Evaluations and diagnostics – Basic Ordinary Least Squares (OLS) regression analysis without considering the clustering, stratification and weighting factors

There are seven classical OLS assumptions for linear regression. The first six are mandatory to produce the best estimates. Assumptions for the OLS regression model include:

(a) The regression model is linear in the coefficients and the error term. The relationships between the predictors and the outcome variable should be linear.
(b) The error term has a population mean of zero.

(c) All independent variables are uncorrelated with the error term.

(d) Observations of the error term are uncorrelated with each other. The errors associated with one observation are not correlated with the errors of any other observations.

(e) The error term has a constant variance (no heteroscedasticity - homogeneity of variance).

(f) No independent variable is a perfect linear function of other explanatory variables.

(g) The error term is normally distributed (optional). The errors should be normally distributed – technically normality is necessary only for hypothesis tests to be valid, estimation of the coefficients only requires that the errors be identically and independently distributed (Wooldridge, 2016).

An OLS analysis was carried out to examine the direct effect of the independent variables (i.e. distance in km to the nearest school, region of residence, household size, urban/rural, dwelling unit type, main material for the wall, source of energy, internet access, sex of respondent, age of respondent, marital status, type of employer, contribution to investment and type of work) on the dependent variable, income, and the results are shown in Table 11. The ordinary multiple regression model for this study's income data contained the following variables with income as a dependent variable and other variables as independent ones displayed in Model 1:

Dependent variable*: income (q10_02_1_log)*

Independent variables*: education (q03_05); distance in km to high school (q02_51_04); region (i.region); types of dwelling unit (i. dwelling_unit_type); the main material used for the walls (q02_03); main source of energy for lighting* *(i.main_source_energy_lighting); internet access (i.internet_access); household members (hh_members); age of respondent (q01_06_y); sex of respondent (i.q01_02); marital status (i.marital_status); type of employer (i. employer_type); contribution to investment (i.q20_04); type of work (i.work_type); employment period (employ_period); total hours worked (totalhrs); type of transport mode (i.transport_mode)*

Table 11: Outputs from the OLS regression analysis without considering the clustering, stratification and weighting factors

Source	SS	df	MS	Number of obs	=	5,556
				F(48, 5507)	=	133.8300
Model	9275.14	48	193.2321	Prob > F	=	<0.0001
Residual	7951.42	5,507	1.4439	R-squared	=	0.5384
				Adj R-squared	=	0.5344
Total	17226.6	5,555	3.1011	Root MSE	=	1.2016
Income	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
Region						
Erongo	-0.2801	0.0731	-3.8300	<0.0001	-0.4234	-0.1368
Hardap	-0.1109	0.0785	-1.4100	0.1580	-0.2648	0.0430
Kavango East	-0.8457	0.1061	-7.9700	<0.0001	-1.0536	-0.6377
Kavango West	-0.6461	0.1116	-5.7900	<0.0001	-0.8649	-0.4272
Khomas	-0.1088	0.0691	-1.5700	0.1160	-0.2443	0.0268
Kunene	-0.1208	0.1115	-1.0800	0.2780	-0.3393	0.0977
Ohangwena	-0.9090	0.1037	-8.7600	<0.0001	-1.1123	-0.7057
Omaheke	-0.5885	0.0934	-6.3000	<0.0001	-0.7716	-0.4054
Omusati	-0.5827	0.1076	-5.4100	<0.0001	-0.7937	-0.3716
Oshana	-0.5175	0.0837	-6.1800	<0.0001	-0.6816	-0.3533
Oshikoto	-0.1552	0.0918	-1.6900	0.0910	-0.3352	0.0248
Otjozondjupa	-0.3107	0.0737	-4.2100	<0.0001	-0.4553	-0.1662
Zambezi	-1.4472	0.1068	-13.5500	<0.0001	-1.6566	-1.2378
//Kharas (reference group)						
Dwelling unit type						

Traditional dwelling	-0.1266	0.0797	-1.5900	0.1120	-0.2828	0.0296
Improvised housing unit	0.1084	0.0635	1.7100	0.0880	-0.0162	0.2330
Town house/Apartment	-0.0123	0.0514	-0.2400	0.8100	-0.1131	0.0885
Others	-0.2364	0.0813	-2.9100	0.0040	-0.3958	-0.0771
Detached house (reference group)						
Main material for the wall						
Corrugated iron/Zinc	-0.3207	0.0599	-5.3500	<0.0001	-0.4381	-0.2033
Sticks, wooden poles, clay and/or cow-d..	-0.2574	0.0785	-3.2800	0.0010	-0.4112	-0.1036
Others	0.0757	0.0708	1.0700	0.2850	-0.0630	0.2144
Cement blocks/Bricks/Stones (reference group)						
Contribution to investment						
No	-0.1597	0.0357	-4.4800	<0.0001	-0.2297	-0.0898
Yes (reference group)						
Age of respondent	0.0091	0.0020	4.4900	<0.0001	0.0051	0.0131
Sex of respondent						
Male	0.3729	0.0346	10.7900	<0.0001	0.3051	0.4406
Female (reference group)						
Household size	-0.0236	0.0061	-3.8700	<0.0001	-0.0355	-0.0116
Marital status						
Married	0.1598	0.0379	4.2100	<0.0001	0.0854	0.2342
Others	0.2451	0.0850	2.8900	0.0040	0.0786	0.4117
Never married (reference group)						
Type of employer						
Government, public sector or army	0.4496	0.0475	9.4700	<0.0001	0.3565	0.5427
Private household/ individual	-0.5624	0.0447	-12.5900	<0.0001	-0.6499	-0.4748
Others	0.0751	0.1270	0.5900	0.5540	-0.1739	0.3241

Private Company, State-owned enterprise, Commercial (reference group)						
Type of work						
Wage/ Salary Activity: Not in agriculture	0.3234	0.0504	6.4100	<0.0001	0.2246	0.4222
Wage/Salary Activity: In Agriculture (reference group)						
Education	0.0468	0.0056	8.3000	<0.0001	0.0357	0.0578
Internet access						
No internet access	-0.2762	0.0379	-7.2900	<0.0001	-0.3506	-0.2019
internet access (reference group)						
Educational attainment						
Primary	0.2291	0.0827	2.7700	0.0060	0.0669	0.3913
Secondary	0.2562	0.0820	3.1200	0.0020	0.0955	0.4169
Tertiary	0.8148	0.0975	8.3600	<0.0001	0.6237	1.0059
Not stated	-0.1144	0.3034	-0.3800	0.7060	-0.7092	0.4803
No formal education (reference group)						
Source of energy for lighting						
Battery lamp/ Torch/ Cell phone	-0.1215	0.0590	-2.0600	0.0400	-0.2372	-0.0058
Candles	-0.2386	0.0613	-3.8900	<0.0001	-0.3588	-0.1184
Solar energy	-0.0150	0.0832	-0.1800	0.8570	-0.1781	0.1481
Others	-0.0580	0.0954	-0.6100	0.5430	-0.2449	0.1289
Electricity from mains (reference group)						
Distance to nearest high school	0.0000	0.0000	4.1700	<0.0001	0.0000	0.0001
Total hours worked						
8+	0.8109	0.1416	5.7300	<0.0001	0.5334	1.0884
less than 8 working hours (reference group)						
Employment period						
1-5 years	0.9077	0.0418	21.7000	<0.0001	0.8257	0.9897
6-10 years	1.0440	0.0584	17.8700	<0.0001	0.9295	1.1586
11 years or more	1.2268	0.0613	20.0100	<0.0001	1.1066	1.3470
Less than 1 year (reference group)						

Transport mode						
Public (taxi/minibus/bus)	0.3659	0.0470	7.7800	<0.0001	0.2737	0.4582
Private (sedan/bakkie)	0.5359	0.0476	11.2500	<0.0001	0.4425	0.6292
Others	0.2344	0.0930	2.5200	0.0120	0.0521	0.4166
Walk (reference group)						
_cons	7.2741	0.2023	35.9600	<0.0001	6.8775	7.6706

Table 11 presents results for Model 1 which was produced without considering the critical elements namely clustering, stratification and weighting factors, which by default meant that the design was from a simple random sampling and standard errors are calculated as such. Almost, all the variables were statistically significant ($p<0.001$) at a 5% level of significance.

Furthermore, the endogeneity of education was also ignored in the model. From the regression output (Table 11), it can be observed that there was a positive relationship between education and (logarithm) income and the linear specification for education indicated a return of 4.7 percent per year. Apart from Hardap, Khomas, Kunene and Oshikoto regions that were not significant in the model, all other regions compared to //Kharas region, had reduced income. Keeping other predictor variables constant, individuals that lived in Ohangwena region in comparison with those who lived in the //Kharas region, had their income reduced by 90.3 percent. Individuals who had tertiary as their highest educational attainment had their income increased by 81.4 percent compared to those with no formal education. For those whose type of dwelling were others as compared to the detached (modern) house, their income had reduced by 23.6 percent.

Moreover, income declined by approximately 16 percent for the households that were not contributing to investment in relation to those that had investment contribution(s). Individuals who engaged in activities that were wage/salary related but not in agriculture as compared to those in the agriculture sector, had income increase of 32.3 percent. For those that did not have access to the internet, their income declined by 27.6 percent. The income for individuals who worked for more than 8 hours a day compared to those who worked less than 8 hours a day increased by 81.1 percent. The employment period also played a role in the amount of income an individual received, regardless of the duration they have been employed as there was an observed positive relationship. For individuals who had been employed for more than 11 years as compared to those who had been employed for less than one year, their income increased by 122.7 percent.

4.2.2 Model 2 Evaluations and diagnostics – Basic Ordinary Least Squares (OLS) regression analysis considering the clustering, stratification and weighting factors

Since the 2015/2016 NHIES was a household-based survey, a two-stage stratified cluster-sampling/complex design survey of which the three elements (clustering, stratification and weighting) should be considered during data analysis.

Table 12 shows the results of the OLS regression model of income and other factors. The model considers the clustering and stratification effect of the survey. Education was treated as an exogenous variable in the model. The confidence intervals were marginally wider and the standard errors were slightly larger when stratification, clustering and weighting were considered. The dwelling unit type, contribution to investment, main source of energy for lighting and distance (km) to high school were insignificant in Model 2. The education and logarithmic income was still positively correlated, and the linear specification for education predicted an annual return of 5.7 percent. In Model 2, the income for individuals who worked for more than 8 hours a day compared to those who worked less than 8 hours a day increased by 90.1 percent. The individuals who were married as compared to those who were never married, had their income increased by approximately 23 percent. For those who lived in the dwelling units' whose wall materials were made of corrugated iron/zinc as compared to those whose dwelling units' wall materials were made of cement blocks/bricks/stones, their income reduced by almost 50 percent.

Essentially, the household surveys utilize complex sample designs to control survey costs. Since this study data was from a complex survey design, Model 2 estimates were slightly different from those obtained by Model 1 (without considering clustering and stratification). The fact that survey data are obtained from units selected with complex sample designs needs to be considered in the survey analysis, in that weights need to be used in analysing survey data and variances of survey estimates need to be computed in a manner that reflects the complex sample design (Chromy, 2004). In contrast to Simple Random Sampling, a complex sample is a probability sample created by utilising sampling techniques including stratification, clustering, and weighting. These techniques are intended to increase statistical efficiency, lower costs, or increase precision for subgroup analysis. Estimating the accuracy of survey estimates necessitates taking into consideration the sample design's complexity (Chromy, 2004). Unbiased estimates with quantifiable sampling error remain a possibility, and the equality of selection probabilities may no longer hold for independent observations that are independently and identically distributed.

Table 12: Outputs from the OLS regression analysis considering the clustering, stratification and weighting factors

Survey: Linear regression						
Number of	strata	=	28	Number of obs	=	5556
Number of	PSUs	=	802	Population size	=	1484572
				Design df	=	774
				F(48, 727)	=	68.200
				Prob > F	=	<0.0001
				R-squared	=	0.5623
Income	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
Region						
Erongo	-0.4222	0.1473	-2.8700	0.0040	-0.7114	-0.1330
Hardap	-0.1985	0.1295	-1.5300	0.1260	-0.4527	0.0558
Kavango East	-0.9466	0.2770	-3.4200	0.0010	-1.4903	-0.4028
Kavango West	-0.7432	0.1725	-4.3100	<0.0001	-1.0817	-0.4047
Khomas	-0.0695	0.0952	-0.7300	0.4650	-0.2563	0.1173
Kunene	-0.1400	0.1262	-1.1100	0.2680	-0.3877	0.1077
Ohangwena	-0.9365	0.1532	-6.1100	<0.0001	-1.2372	-0.6357
Omaheke	-0.6973	0.1926	-3.6200	<0.0001	-1.0755	-0.3191
Omusati	-0.6412	0.1683	-3.8100	<0.0001	-0.9717	-0.3107
Oshana	-0.6328	0.1554	-4.0700	<0.0001	-0.9378	-0.3278
Oshikoto	-0.0113	0.1362	-0.0800	0.9340	-0.2786	0.2561
Otjozondjupa	-0.2632	0.0995	-2.6400	0.0080	-0.4585	-0.0678
Zambezi	-1.5469	0.2102	-7.3600	<0.0001	-1.9595	-1.1342
//Kharas (reference group)						
Dwelling unit type						
Traditional dwelling	-0.2505	0.1439	-1.7400	0.0820	-0.5330	0.0320
Improvised housing unit	0.1450	0.1198	1.2100	0.2270	-0.0903	0.3802
Town house/Apartment	-0.1071	0.0884	-1.2100	0.2260	-0.2807	0.0665
Others	-0.2990	0.2417	-1.2400	0.2160	-0.7734	0.1753
Detached house (reference group)						
Main material for the wall						

Corrugated iron/Zinc	-0.4987	0.1306	-3.8200	<0.0001	-0.7550	-0.2423
Sticks, wooden poles, clay and/or cow-dung	-0.2670	0.1412	-1.8900	0.0590	-0.5441	0.0101
Others	0.1876	0.1193	1.5700	0.1160	-0.0466	0.4217
Cement blocks/Bricks/Stones (reference group)						
Contribution to investment						
No	-0.0954	0.0605	-1.5800	0.1150	-0.2141	0.0233
Yes (reference group)						
Age of respondent	0.0085	0.0033	2.5600	0.0110	0.0020	0.0150
Sex of respondent						
Male	0.3437	0.0456	7.5400	<0.0001	0.2542	0.4332
Female (reference group)						
Household size	-0.0348	0.0125	-2.7800	0.0060	-0.0595	-0.0102
Marital status						
Married	0.2286	0.0550	4.1600	<0.0001	0.1207	0.3366
Others	0.2484	0.1209	2.0500	0.0400	0.0110	0.4857
Never married (reference group)						
Type of employer						
Government, public sector or army	0.5052	0.0705	7.1600	<0.0001	0.3668	0.6437
Private household/ individual	-0.4195	0.0727	-5.7700	<0.0001	-0.5622	-0.2767
Others	-0.0913	0.2042	-0.4500	0.6550	-0.4922	0.3096
Private Company, State-owned enterprise, Commercial (reference group)						
Type of work						
Wage/ Salary Activity: Not in agriculture	0.3716	0.0870	4.2700	<0.0001	0.2007	0.5424
Wage/Salary Activity: In Agriculture (reference group)						
Education	0.0569	0.0103	5.5100	<0.0001	0.0366	0.0772

Internet access						
No internet access	-0.2824	0.0708	-3.9900	<0.0001	-0.4214	-0.1434
internet access (reference group)						
Educational attainment						
Primary	0.3092	0.1390	2.2300	0.0260	0.0364	0.5821
Secondary	0.2062	0.1427	1.4500	0.1490	-0.0738	0.4863
Tertiary	0.6645	0.1713	3.8800	<0.0001	0.3282	1.0009
Not stated	-0.5775	0.5301	-1.0900	0.2760	-1.6181	0.4631
No formal education (reference group)						
Source of energy for lighting						
Battery lamp/ Torch/ Cell phone	-0.0395	0.1162	-0.3400	0.7340	-0.2675	0.1886
Candles	-0.1476	0.1255	-1.1800	0.2400	-0.3940	0.0988
Solar energy	0.0638	0.1391	0.4600	0.6460	-0.2093	0.3370
Others	0.1769	0.1847	0.9600	0.3380	-0.1857	0.5395
Electricity from mains (reference group)						
Distance to nearest high school	0.0000	0.0000	1.8500	0.0650	0.0000	0.0001
Total hours worked						
8+	0.9011	0.3919	2.3000	0.0220	0.1318	1.6705
less than 8 working hours (reference group)						
Employment period						
1-5 years	0.9710	0.0673	14.4200	<0.0001	0.8389	1.1032
6-10 years	1.1405	0.0948	12.0300	<0.0001	0.9544	1.3266
11 years or more	1.2773	0.1003	12.7300	<0.0001	1.0804	1.4743
Less than 1 year (reference group)						
Transport mode						
Public (taxi/minibus/bus)	0.3517	0.0714	4.9300	<0.0001	0.2116	0.4918
Private (sedan/bakkie)	0.5746	0.0798	7.2000	<0.0001	0.4179	0.7313
Others	0.2549	0.1202	2.1200	0.0340	0.0190	0.4908
Walk (reference group)						
_cons	7.0444	0.4301	16.3800	<0.0001	6.2001	7.8886

4.2.3 Model 3 Evaluations and diagnostics – Instrumental Variables & Two Stage Least Squares (IV-2SLS) regression analysis

Despite considering the three variables (weighting, clustering, and stratification) in Model 2, the endogeneity problem has not been considered in both Model 1 and Model 2. The net effect of clustering, stratification and weighting, therefore, is that the standard errors of these complex sample designs tend to be different (smaller or larger, but usually larger) than those of a simple random sample. Design effect (DEFF) refers to the variation in the accuracy of the estimates generated by a complex design versus a simple random sample. In Model 3 (Table 13), considering stratification, clustering and weighting factors, the confidence interval produced by the 2SLS model were slightly wider and standard errors were larger than those from the OLS model (Model 1).

In this Model 3, results presented in Table 13, education was treated as an endogenous variable being influenced by internet access, educational attainment, main source of energy for lighting and distance to the nearest high school. It can be observed that there was a positive relationship between education and logarithmic income. Using the two-stage least regression, a unit increase in the age increased the salary/income by approximately 1.25 percent as opposed to 0.91 percent in the OLS model (Model 1). In spite of that, treating education endogenous, the relationship between logarithmic income and education was such that an additional year of education contributed to about 13 percent of the income. Meanwhile, an additional household member reduced the income by approximately 4 percent. Those who were married as compared to those who were never married, had their income increased by 24 percent. Those working for more than 8 hours a day, compared to those who did not, had their income increased by 84 percent. Furthermore, being a resident in the Ohangwena region compared to being a resident in the //Kharas region decreased individual income by about 98 percent.

With respect to sex, with reference to female and holding other variables constant, being a male increased individual income with 37 percent. Being an employee who received wage/salary not in agriculture (meaning from activities that are off-farm) as compared to those who received wage/salary in agriculture, resulted in an income increment by 28.2 percent. Taking the main building material for the wall into consideration, being in a house built with corrugated iron and sticks, wooden poles, clay and/or cow-dung as referenced to the detached/modern houses, reduced the income by 57 percent and 29 percent respectively.

Table 13: Outputs from the IV-2SLS regression analysis

Survey: Linear regression						
Number of	strata	=	28	Number of obs	=	5556
Number of	PSUs	=	802	Population size	=	1484572
				Design df	=	774.000
				F(38, 737)	=	66.340
				Prob > F	=	<0.0001
				R-squared	=	0.5342
Income	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
Education	0.1270	0.0243	5.2300	<0.0001	0.0794	0.1747
Region						
Erongo	-0.5138	0.1444	-3.5600	<0.0001	-0.7973	-0.2303
Hardap	-0.1754	0.1290	-1.3600	0.1740	-0.4286	0.0777
Kavango East	-0.8894	0.2717	-3.2700	0.0010	-1.4227	-0.3561
Kavango West	-0.7111	0.1714	-4.1500	<0.0001	-1.0476	-0.3747
Khomas	-0.0192	0.0975	-0.2000	0.8440	-0.2106	0.1722
Kunene	-0.2346	0.1155	-2.0300	0.0430	-0.4613	-0.0079
Ohangwena	-0.9772	0.1633	-5.9800	<0.0001	-1.2978	-0.6566
Omaheke	-0.7164	0.2205	-3.2500	0.0010	-1.1492	-0.2835

Omusati	-0.7479	0.1763	-4.2400	<0.0001	-1.0940	-0.4018
Oshana	-0.7501	0.1524	-4.9200	<0.0001	-1.0493	-0.4510
Oshikoto	-0.1421	0.1416	-1.0000	0.3160	-0.4201	0.1358
Otjozondjupa	-0.2433	0.1063	-2.2900	0.0220	-0.4521	-0.0346
Zambezi	-1.6696	0.2174	-7.6800	<0.0001	-2.0964	-1.2427
//Kharas (reference group)						
Dwelling unit type						
Traditional dwelling	-0.3566	0.1388	-2.5700	0.0100	-0.6290	-0.0841
Improvised housing unit	0.0926	0.1263	0.7300	0.4640	-0.1553	0.3406
Town house/Apartment	-0.1274	0.0932	-1.3700	0.1720	-0.3104	0.0556
Others	-0.6461	0.2784	-2.3200	0.0210	-1.1926	-0.0997
Detached house (reference group)						
Main material for the wall						
Corrugated iron/Zinc	-0.5697	0.1326	-4.3000	<0.0001	-0.8300	-0.3094
Sticks, wooden poles, clay and/or cow-dung	-0.2879	0.1506	-1.9100	0.0560	-0.5836	0.0077
Others	0.1909	0.1147	1.6600	0.0960	-0.0342	0.4160
Cement blocks/Bricks/Stones (reference group)						
Contribution to investment						
No	-0.1079	0.0635	-1.7000	0.0900	-0.2326	0.0168
Yes (reference group)						
Age of respondent	0.0125	0.0037	3.4100	0.0010	0.0053	0.0197
Sex of respondent						
Male	0.3700	0.0476	7.7700	<0.0001	0.2765	0.4635
Female (reference group)						
Household size	-0.0361	0.0136	-2.6500	0.0080	-0.0629	-0.0094
Marital status						
Married	0.2386	0.0574	4.1600	<0.0001	0.1260	0.3513

Others	0.2408	0.1230	1.9600	0.0510	-0.0007	0.4823
Never married (reference group)						
Type of employer						
Government, public sector or army	0.4825	0.0827	5.8300	<0.0001	0.3201	0.6450
Private household/ individual	-0.4122	0.0774	-5.3300	<0.0001	-0.5641	-0.2604
Others	-0.1466	0.2246	-0.6500	0.5140	-0.5875	0.2943
Private Company, State-owned enterprise, Commercial (reference group)						
Type of work						
Wage/ Salary Activity: Not in agriculture	0.2815	0.0951	2.9600	0.0030	0.0948	0.4682
Others	0.0000	(empty)				
Wage/Salary Activity: In Agriculture (reference group)						
Total hours worked						
8+	0.8352	0.3768	2.2200	0.0270	0.0955	1.5749
Less than 8 working hours (reference group)						
Employment period						
1-5 years	0.9767	0.0716	13.6400	<0.0001	0.8361	1.1173
6-10 years	1.1536	0.0949	12.1600	<0.0001	0.9673	1.3398
11 years or more	1.2864	0.1042	12.3500	<0.0001	1.0820	1.4909
Less than 1 year (reference group)						
Transport mode						
Public (taxi/minibus/bus)	0.3338	0.0772	4.3300	<0.0001	0.1823	0.4852
Private (sedan/bakkie)	0.5411	0.0861	6.2900	<0.0001	0.3721	0.7100
Others	0.2263	0.1235	1.8300	0.0670	-0.0161	0.4688
Walk (reference group)						
_cons	6.5162	0.4927	13.2200	<0.0001	5.5490	7.4835

4.2.4 The OLS and 2SLS model coefficients

The OLS and 2SLS model analysis results were further compared together (Table 18 in Appendix II). Both OLS and 2SLS estimates were consistent if all the variables were exogenous. If 2SLS and OLS differ significantly, then it can be concluded that education must be endogenous (i.e., maintaining that the factors affecting education were exogenous). This sentiment can be observed in Table 18.

4.2.5 Postestimation tests: (IV-2SLS) regression

A series of postestimation tests were performed while implementing IV-based estimations. These tests assumed that the error terms were independently and identically distributed. Instruments are valid if they affect income through the instrument equation.

(a) Endogeneity Test (Durbin and Wu-Hausman tests)

The Durbin score and Wu-Hausman p-values were significant ($p<0.001$). This implied that the null hypothesis (not endogenous) was rejected and, thus, the education variable was endogenous. For this reason, it can be concluded that there was an endogeneity problem, and as such, the OLS results cannot be relied on. Hence, other modelling alternatives using IV needed to be explored. The chosen IV must be highly correlated with the endogenous variables but should be uncorrelated with the disturbance term (error term). All the identified IV considered in this study (internet access, highest level of education, main source of lighting for lighting and distance in km to high school) were statistically significant determinants of years of education, and the regression model was statistically significant at a 5% level of significance with $p<0.001$, $F(1,5516) = 72.71$ and $\chi^2 = 72.29$ from the Durbin and Wu-Hausman test.

(b) Testing for weak instruments (F-test)

Instrument quality affects how well endogeneity tests work. In other words, results from weak or endogenous instruments are questionable, but results from stronger or exogenous tools show endogeneity. In light of this, tests that relied on subpar and/or endogenous tools may provide false positives. From Table 15, testing for the weak instruments had revealed that there was no-weak instrument and the IV were valid. All of the identifying IV considered in this study (internet access, highest level of education, main source of lighting for lighting and distance in km to high school) were statistically significant determinants of years of education. Furthermore, the number of IV was more than the number of endogenous variables, therefore, a test needed to be performed to test whether the instruments were correlated with the error term of the system. If they were, then they were not good instruments to use.

Table 14: Testing for weak instruments (F-Statistics)

First-stage regression summary statistics					
Ho: Instruments are weak Ha: Instruments are not weak					
Variable	R-square	Adjusted R-square	Partial R-square	F(10,5508)	Prob > F
Education	0.4313	0.4265	0.2089	145.464	<0.0001
Minimum eigenvalue statistic			145.464		
Critical Values	# of endogenous regressors:			1	
Ho: Instruments are weak	# of excluded instruments:			10	
	5%	10%	20%	30%	
2SLS relative bias	20.74	11.49	6.61	4.86	
	10%	15%	20%	25%	
2SLS Size of nominal 5% Wald test	38.54	20.88	14.78	11.65	
LIML Size of nominal 5% Wald test	3.68	2.84	2.46	2.25	

The test being conducted (Table 15) was a two-stage least squares model. The test statistic of 145.46 exceeded its critical value of 38.54. Thus, there was no weak-instrument problem. Furthermore, instruments are likely to be strong (relevant) if the *F*-statistic is greater than 9.08 with three instruments and greater than 10.83 with five instruments (Ullah et al., 2018). Since this model had only one endogenous regressor, the minimum eigenvalue statistic (145.464) was equal to the *F* statistic obtained (145.464). In addition, since the F-value (145.464) was greater than any of the critical values in the table (Table 15), it can be concluded that the IV were not weak but were good instruments.

(c) Instrument Validity test

A test of over-identifying restrictions to verify the validity of the excluded instruments was further performed and the obtained result shown in Table 16. If there was just one instrument z_j for each endogenous variable y_j, then the model can be said to be "just-identified" and tests whether z_j is uncorrelated with the error u. However, if the model had multiple instruments for at least one endogenous variable, then model can be said to be "over-identified" and tests whether the over-identifying instruments were good instruments can be performed (Li et al., 2021). A model is "just-identified" when the number of instruments (i.e., IV) equals the number of endogenous variables, and a model is "over-identified" when the number of instruments is greater than the number of endogenous variables. In this study, the number of instruments was four (internet access, highest level of education, main source of energy for lighting and distance in km to high school) while the number of endogenous variables was just one (education). Thus, since the number of instruments (4) was greater than the number of endogenous variables (1) and p<0.001, the IV model can be said to be over-identified.

Table 15: Validity of instruments test

Tests of over-identifying restrictions:
Sargan (score) chi2(9) = 161.635 (p <0.001)
Basmann chi2(9) = 165.04 (p <0.001)

(d) Test of heteroskedasticity for IV estimation

The issue of heteroskedasticity occurs in the error term. Heteroskedasticity implies that the variance of the error term is non-constant (i.e., it increases or decreases with changes in the variables). The test of heteroskedasticity for IV estimation performed was the Breusch-Pagan / Cook-Weisberg test for heteroskedasticity and the result shown in Table 17. A small p-value indicates that residual variance was non-constant (heteroscedastic). From Table 17, since the p-value (p<0.0001) is small, thus, it can be concluded that residual variance was non-constant (heteroscedastic).

Table 16: Test for heteroskedasticity

H_0: Constant variance H_1: Non-constant variance
F(48 , 5507) = 2.7900
(Prob > F) <0.0001

4.2.6 Best fit modelling strategy

Three different modelling strategies were carried out, namely: (i) Basic OLS Regression Analysis without considering the clustering, stratification and weighting factors, (ii) Basic OLS Regression Analysis considering the clustering, stratification and weighting factors, and (iii) Instrumental Variable & Two Stage Linear Regression (IV-2SLS). The best fit modelling strategy to use for exploring the relationship between income and education, and other socio-economic and socio-demographic factors was determined using the Durbin and Wu-Hausman test of endogeneity (see section 4.2.5). In addition, the parameter estimates for the best model were carried out and the interpretation of the results. From the results obtained from the Durbin and Wu-Hausman test of endogeneity (Table 14, the presence of endogeneity was confirmed. Thus, education must be endogenous, making the OLS strategies unreliable to use. For this reason, the IV-2SLS regression analysis was identified as the best modelling strategy to use and the obtained results shown in Table 13.

From Table 13 at a 5% level of significance, it can be concluded that income was negatively affected within individuals in the Erongo (β = -0.5138, p<0.0001, 95% CI:-0.7973 to -0.2303), Kavango East (β = -0.8894, p<0.0001, 95% CI: -1.4227 to -0.3561), Kavango West (β = -0.7111, p<0.0001, 95% CI: -1.0476 to -0.3747), Kunene

(β= -0.2346, p<0.0001, 95% CI-0.4613 to -0.0079), Ohangwena (β= -0.9772, p<0.0001, 95% CI: -1.2978 to -0.6566), Omaheke (β= -0.7164, p=0.0010, 95% CI: -1.1492 to -0.2835), Omusati (β= -0.7479, p<0.0001, 95% CI: -1.0940 to -0.4018), Oshana (β= -0.7501, p<0.0001, 95% CI: -1.0493 to -0.4510), Otjozondjupa (β= -0.2433, p=0.0050, 95% CI: -0.4521 to -0.0346) and Zambezi (β= -1.6696, p<0.0001, 95% CI: -2.0964 to -1.2427) regions, compared to those in the //Karas region. Similarly, income was undesirably affected by the household size (β= -0.0361, p<0.0001, 95% CI: -0.0629 to -0.0094) and types of dwelling units; households dwell in traditional set up (β = -0.3566, p<0.0001, 95% CI: -0.6290 to -0.0841), others (β = -0.6461, p<0.0001, 95% CI: -1.1926 to -0.0997) as compared to those whose dwelling units type was detached house. Additionally, for individuals whose household's main material for the wall was corrugated iron/zinc (β = -0.5697, p<0.0001, 95% CI: -0.8300 to -0.3094) as associated to cement blocks/bricks/stones had their income negatively affected.

Furthermore, the age of the respondent (β = 0.0125, p<0.0001, 95% CI: 0.0053 to 0.0197) and males sex (β = 0.3700, p<0.0001, 95% CI: 0.2765 to 0.4635) as compared to females positively affected their income. Further to this, the income by married couples (β = 0.2386, p<0.0001, 95% CI: 0.1260 to 0.3513) and others (β = 0.2408, p<0.0001, 95% CI: -0.0007 to 0.4823) were positively affected as compared to never married individuals. The employees who worked for the government, public sector or army (β = 0.4825, p<0.0001, 95% CI: 0.3201 to 0.6450) and those who had worked for the private household/individual (β = -0.4122, p<0.0001, 95% CI: -0.5641 to -0.2604) had their income positively and negatively affected respectively when comparing them to those who work in private company, state-owned enterprise, commercial employers. Salaried individuals who worked in off-farm (not-in agriculture) industry (β = 0.2815, p<0.0001, 95% CI: 0.0948 to 0.4682) had their income positively affected compared to the salaried individuals who worked in the agricultural industry (Table 13).

Moreover, income was also positively influenced by the individuals who worked for more than 8 hours per day (β =0.8352, p<0.0001, 95% CI: 0.0955 to 1.5749) as compared to those who worked less than 8 hours a day. Consequently, individuals who had worked for 1 to 5 years (β = 0.9767, p<0.0001, 95% CI: 0.8361 to 1.1173); 6 to 10 years (β = 1.1536, p<0.0001, 95% CI: 0.9673 to 1.3398) and 11 years or more (β = 1.2864, p<0.0001, 95% CI: 1.0820 to 1.4909) had their income positively affected as compared to those who had been employed for less than a year. Similarly, with reference to walking, individuals who took public (taxi/minibus/bus) (β = 0.3338, p<0.0001, 95% CI: 0.1823 to 0.4852) and private (sedan/bakkie) (β = 0.5411, p<0.0001, 95% CI: 0.3721 to 0.7100) mode of transportation to work had their income affected in a positive way.

CHAPTER 5: DISCUSSION, CONCLUSION AND RECOMMENDATIONS

Chapter 5 summarises the key research findings and conclusions of this study, as well as proposes some key recommendations brought forward from this study's findings. The Chapter also presents the limitation and delimitation of this study.

5.1 Discussions

This study of endogeneity and instrumental variables of education on income in Namibia highlighted the importance of education as an endogenous variable of income. Treating education as an exogeneous variable impacts the model. It is important to consider the correct methodology/tools to deal with this issue. These findings are pertinent at this time when the country is still grappling with a high unemployment rate that could be checked by increased educational attainment as a determinant of income. The number of years that children spend in school is also influenced by the parents' financial capabilities. Additionally demonstrated to have substantial effects on years of schooling are household size and place of residence.

The results indicated that using the IV estimators instead of OLS, the return to education increased to 13% when education was treated as endogenous. This estimate was higher compared to the OLS estimate of 5%. The study findings are similar to those of (Parker & van Praag, 2006) in their article of "Schooling, Capital Constraints and Entrepreneurial Performance" which confirmed that between IV and OLS conducted for employees, the IV estimate is substantially higher than the OLS estimate, 13.7% versus 7.2%. Parker and van Praag (2006) employed Instrumental Variable (IV) techniques in their studies to estimate the effect of education (formal schooling) on entrepreneurial income. The findings were that education's effect on entrepreneurial income increases from 7.2% in the ordinary least squares (OLS) model to a 13.7% reward of additional income per extra year of education in the IV model (an increase of about 90%). This is also confirmed by Coady & Dizioli (2015) that the average number of years a person attends school is positively correlated with income disparity, which is consistent with rising or steadily growing returns to more education. Block et al. (2012) demonstrated that income and education are endogenous variables and ignoring such in the model would lead to erroneous conclusions. Moreover, the errors are propagated in other factors, if the endogenous variable is erroneously omitted (Vo et al., 2019).

Further, the findings of this study resonated with other works presented on the causal relationship between education and earnings as stipulated by Lv (2017). The estimated returns to education are consistently found to be 20–40% higher than the corresponding OLS estimates in research using instrumental variables that are based on institutional changes in the educational system. This conclusion might be explained in part by the fact that some subgroups' marginal returns to education are larger than the average marginal return to education for the population as a whole, particularly for comparatively disadvantaged groups with poor educational performance. Similarly, Chevalier et al. (2013) established that the IV-point estimate for daughters, at 0.07, is ten times larger than the OLS estimate. The influence of paternal schooling on sons is never significant and fluctuates greatly.

5.1.1 Region

All regions were statistically significant in the model except Hardap, Khomas, Kunene and Oshikoto. There are regions mainly rural and are dominated by agricultural activities (subsistence and commercial) and some are a mixture of both (off-farm and on-farm). The north central and northeast regions, namely Kavango East, Kavango Ohangwena, Omusati, Oshana and Oshikoto are mainly dominated by subsistence farming of which sales are very minimal, hence their contributions are inversely related to income (salaries and wages). Additionally, the mean incomes for these regions are the lowest when comparing to the mean income of //Karas region (Table 3). //Kharas is one of the regions where most of the off-farm activities take place e.g mining. The regions dominated by most diversified economic activities (commercial farming, mining). Unlike Khomas where most of the people are regular salary/wage earners who work for the government or other organisations like parastatal corporations and business. Being the most important country's most important metropolis, it has drawn the attention of rural residents looking for employment. The findings of the study have confirmed that receiving salaries/wages by performing activities that are off-farm nature as compared to those who do activities in the agriculture industry increases income.

In this study, the endogenous variable under discussion is education which is measured in terms of the total number of years of schooling one has schooled including repetitions. Place of residence/region remains an important factor in the determination of education. The place of residence is influenced by the parents' educational background. Marbuah, 2016's findings confirm that parents with higher levels of education are more likely to live in regions with urban areas, while parents with lower levels of education are more likely to live in rural areas. This indicates that parents with advanced education levels give their kids access to a good infrastructure and social facilities in a convenient setting. The funds and endowments of urban schools are likely not available to rural schools. This is seen in both the teaching "skill" they draw and any equipment the rural schools may have. There might be fewer high-tech labs and opportunities for simulation. It is possible that rural schools do not have the partnerships necessary to support career-guidance programmes, internship and study abroad programmes.

5.1.2 Household size/members

The inverse relationship between income and household members as demonstrated by the study findings clearly verify the impact of large households. Rural areas in Namibia are associated with large household sizes that have a negative influence on children's years of schooling, while their urban counterparts generally have smaller household sizes and with a positive influence established. The findings have been confirmed elsewhere by (Marbuah, 2016) who emphasized that large household sizes are also linked to parents with more children as their level of education is also lower as compared to their counterparts who have attained the highest level of education. In most cases, such low-educated parents put subsistence farming before their children's education. Namibia is a youthful nation; an additional household member could mostly be children and do not add any income in the form of salaries/wages. Since this study only looked at the salary/wage as in income, one could look at other sources of income and it could depict a different relationship between income and household size. There are children, especially the vulnerable children who receives monthly grants from government and for some households, this is the main source of livelihoods. Furthermore, workers below management qualify to work overtime for most employers and a picture might be different with the consideration of income in the form of overtime.

5.1.3 Marital status

From this study, the marital status of respondent was statistically significant and is positively related to income. The findings reason with Chowa et al. (2012) who indicate that compared to single parents, married parents are more likely to be involved in their children's education In addition, the combined income of married couples can provide a greater chance of building wealth which ultimately accord them an opportunity to send their children to better schools. They may experience individual net worth increases over their singletons in the same age group. They may have more financial accountability as it pressures each couple to put their personal financial act together due to increased responsibilities as opposed to those never been married. If one member of the couple loses the job, the household might still continue functioning although at a constrained level as opposed to when a single person loses the job. Married couples have a better chance of getting a good mortgage deal because of their greater combined income, consistent payment history, and higher credit. Moreover, due to the fact that married couples' combined income qualifies them for a larger mortgage loan at an ideal interest rate (provided that both spouses have strong credit ratings), mortgage lenders are drawn to these couples. Most of the key components of a household's financial checklist for married couples—including bank accounts, retirement savings, and taxes, among other categories—tend to get significantly better and they build wealth (financial assets and house assets).

5.1.4 Age

Additionally, age plays an important role in demining someone’s income. Those aged older tend to have vast work experience and their probability of working for a well-paid job is higher. This finding agrees with Ozhamaratli et al. (2022) who indicated that the primary economic productive period is between the ages of 25 and 55, and income distribution is very wide throughout this time. Together, all the variables and chance shocks produce a dispersed but reliable distribution. While the income of the 36-45 and 46-55 age groups marginally declined, the relative income share of the 26-35 age group somewhat grew. Mean income had a positive relationship with age until 55 years old for the entire period.

5.1.5 Dwelling unit type and main material for the wall

Findings from this study also reveal that the type of dwelling and main materials for the wall households live in are significantly and negatively associated with income. As evidence by Tables 8 and 9, the mean income for those in town houses or apartments and whose houses are made from concrete is higher than any other group income. The perception of this result implies that some households own apartments/housing units in various places as they could afford to buy them and henceforth, they rent them and this provided for more household income. Prior studies such as Jiburum et al. (2021) have indicated poorer households are more likely to reside in lower-quality housing than wealthier households in the majority of countries, housing quality measurements are particularly relevant at the bottom end of the income distribution. This could be so because with their meagre income they cannot afford to acquire assets such as modern houses.

5.2 Conclusion

Education creates better opportunities in ensuring the competitive advantages of societies and countries at large even at the global level. Consequently, education plays a decisive role in economic performance. Those in society with more education earn higher salaries over their lifetime as well as contribute more to taxes. In addition, education not only raises the level of income, but it also expands new opportunities and create better prospects. One of the human development index elements is education as a measure of quality of life and socio-economic development and this is what determines whether the country is developed, developing or underdeveloped country. Those that manage to attain highest educational levels become better-skilled workers so they can enhance their expertise domain and strive for higher positions within their current employment or elsewhere.

The various stakeholders in education and employment sectors can play a big role in leveraging various population groups. The education actors need to produce labour force with requisite skills and knowledge to be consumed by the market. Moreover, the economy needs various employment opportunities and a conducive environment (accessible business finance, flexible requirements) when it comes to opening business opportunities for small and medium enterprises to operate from. Improved skills through education especially in this time and era of COVID-19 impact is critical particularly in the 4th Industrial Revolution, more jobs will become redundant/obsolete through automation. Education and training become the driving force of the economy to better people’s livelihoods.

The relationship between education and income disparity has been examined by several economists and scholars, who have arrived at the opposite conclusion: there is an inverse association between inequality and education levels (Rehman & Rashidi, 2017). This reflects with Kirima (2019) who stipulated that the educational outcomes of children are influenced by household income as well. Children from high-income households are projected to complete more years of education than those from low-earning households. Moreover, high-income families can afford to pay for extra tutoring, stationery, and books to help their kids perform better in school and advance their education. In contrast, a family with limited financial resources would struggle to cover even the most basic expenses and so be unable to provide additional funding for the education of their children.

The overall objective of this study was to explore different modelling strategies to capture the relationship between income and education and other socio-economic and socio-demographic factors. In addition, also to account for the possible presence of endogeneity of education. The study, henceforth, has confirmed that education is indeed an endogeneous variable and should be handled as such to explain the effects of observed confounders.

5.3 Recommendations

It is hoped that this study will help education/development planners, policy makers to identifying specific areas with educational gaps to design, evaluate programmes and develop strategies aiming at improving the education and reduce income inequality of Namibians. This study's analysis leads us to the following recommendations:

- The need to boost funding for education - Budget cuts are undermining public education. Education requires investment, hence improved funding will improve outcomes. Policy makers should consider investing rather than cutting education funding. Not only that, but governments should also provide more fair education opportunities between urban and rural to level off the disparity of resources and provide subsidies in order to cut down the inequality of intergenerational transmission.

- As part of poverty alleviation strategies, the government should ensure the distribution of equitable resources particularly the educational resources to enhance the development of rural-and underdeveloped regions so that they too can benefit from high-quality education opportunities. The strengthening of provision of state-grants to the vulnerable segments of the population in order to support low-income family. The low-income family can provide a better condition for their children's education if their income can be improved.
- The schools should get background information for the disadvantaged families such as parents with low education levels or lack of interest in education affairs. Teachers need to give special care to the learners from such families. The solid relationship between schools and disadvantaged families can enhance the educational levels of such learners as part of imparting knowledge and skills. This will create a system whereby it will produce better outcomes and strengthen the education opportunities for all.
- The government should through the National Development Plan 6, Poverty Reduction Strategy, Harambee Prosperity Plan, Africa Agenda 2063, Sustainable Development Goals and other development plans/programmes focus on the issues of machine learning and artificial intelligence as part of the educational curriculum. Moreover, if the country is to achieve sustainable development goals, national development goals such as Vision 2030 and National Development Programme 6 (NDP6), then efforts should be made to support adults especially those in the

productive age group with low educational attainment while at the same time considering the impacts of socio-economic factors.

Therefore, this study advocates for areas for further research as follows:

- Searching for valid instruments is a challenging task in estimating IV models that account for regressor-error problems. In addition, the identification of regressor-error correlation can be impossible without such valid instruments. There is an instrument-free approach (Gaussian copula) which presents a practical solution to this circular problem. The instrument-free method can be used to estimate regression parameters and test for regressor-error correlations without the necessity of first finding 'valid' instruments. The possible intervention could be approaching the endogeneity problem using the Latent Instrumental Variables (LIV) to estimate regression parameters consistently regardless of the presence of regressor-error correlations. Testing for endogeneity using the Hausman-LIV test and Generalized Method of Moments (GMM).
- One could further study the impact of education on income using income not transformed (log) and use other models such as the inverse Gaussian distribution to deal with heavy-tailed data such as income. Moreover, since the Government of the Republic of Namibia gives various grants (maintenance grants, special maintenance grant, foster care grant, disability grant and old age grant), one could consider all these grants in the model and observe how they impact the household income. Since there are varying numbers of persons in each family, the study's findings would paint a different image if we additionally looked at the average monthly income of households.

REFERENCES

Abdallah, W., Goergen, M., & Sullivan, N. O. (2015). *Endogeneity: How Failure to Correct for it can Cause Wrong Inferences and Some Remedies*. *26*, 791–804. https://doi.org/10.1111/1467-8551.12113

Afifi, A. A., Kotlerman, J. B., Ettner, S. L., & Cowan, M. (2007). Methods for Improving Regression Analysis for Skewed Continuous or Counted Responses. *Annual Review of Public Health*, *28*(1), 95–111. https://doi.org/10.1146/annurev.publhealth.28.082206.094100

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. *The Oxford Handbook of Leadership and Organizations*, *1*, 93–117. https://doi.org/10.1093/oxfordhb/9780199755615.013.007

Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization*, *6*(3), 285–327. https://doi.org/10.1177/1476127008094339

Belfield, C. R., & Levin, H. M. (2007). The Education Attainment Gap: Who's Affected, How Much, and Why It Matters. *The Price We Pay: Economic and Social Consequences of Inadequate Education*, *4*(2), 1–35. https://doi.org/10.4135/9781412957403.n55

Bernanke, B. S. (2020). Monetary Policy in a New Era. *Evolution or Revolution?*, 3–48. https://doi.org/10.7551/mitpress/11734.003.0004

Bliese, P. D., Schepker, D. J., Essman, S. M., & Ployhart, R. E. (2020). Bridging Methodological Divides Between Macro- and Microresearch: Endogeneity and

Methods for Panel Data. *Journal of Management*, *46*(1), 70–99. https://doi.org/10.1177/0149206319868016

Block, J. H., Hoogerheide, L. F., & Thurik, R. (2012). Are Education and Entrepreneurial Income Endogenous? A Bayesian Analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1570725

Bun, M. J. G., & Harrison, T. D. (2019). OLS and IV estimation of regression models including endogenous interaction terms. *Econometric Reviews, 38*(7), 814–827. https://doi.org/10.1080/07474938.2018.1427486

Chevalier, A., Harmon, C., O' Sullivan, V., & Walker, I. (2013). The impact of parental income and education on the schooling of their children. *IZA Journal of Labor Economics*, *2*(1), 1–22. https://doi.org/10.1186/2193-8997-2-8

Chowa, G., Ansong, D., & Osei-Akoto, I. (2012). *Parental Involvement and Academic Performance in Ghana*.

Chromy, J. R., & Abeyasekera, S.(2004). Statistical analysis of survey data. In. S.I. Santiago, Household Sample Surveys in Developing and Transition Countries, (pp. 389 -417). United Nations Publication

Coady, D., & Dizioli, A. (2018). Income inequality and education revisited: persistence, endogeneity and heterogeneity. Applied Economics, 50(25), 2747–2761. https://doi.org/10.1080/00036846.2017.1406659

Dabla-Norris, E., Kochhar, K., Suphaphiphat, N., Ricka, F., & Tsounta, E. (2015). Causes and Consequences of Income Inequality: A Global Perspective. *Staff Discussion Notes*, *15*(13), 1. https://doi.org/10.5089/9781513555188.006

Doumbia, D., & Kinda, T. (2019). Reallocating Public Spending to Reduce Income Inequality. *IMF Working Papers*, *2019*(188). https://doi.org/10.5089/9781513511863.001

Gretz, R. T., & Malshe, A. (2019). Rejoinder to "Endogeneity bias in marketing research: Problem, causes and remedies." *Industrial Marketing Management*, *77*, 57–62. https://doi.org/10.1016/j.indmarman.2019.02.008

Guisan, A., Edward, T. C., Jr, & Hastie, T. (2016). Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling, 157(157), 89–100. www.elsevier.com/locate/ecolmodel

Hensher, D. A., Johnson, L. W., Hensher, D. A., Johnson, L. W., Louviere, J. J., & Horowitz, J. (2018). *Applied Discrete-Choice Modelling*. https://doi.org/10.4324/9781351140768-8

Hill, A. D., Johnson, S. G., Greco, L. M., O'Boyle, E. H., & Walter, S. L. (2021). Endogeneity: A Review and Agenda for the Methodology-Practice Divide Affecting Micro and Macro Research. *Journal of Management*, 47(1), 105–143. https://doi.org/10.1177/0149206320960533

International Monetary Fund. (2017). *Fostering Inclusive Growth: Group of Twenty (G20) Leader's Summit, Hamburg, Germany*. Retrieved from https://www.imf.org/external/np/g20/pdf/2017/062617.pdf

Jalil, A., & Idrees, M. (2013). Modeling the impact of education on the economic growth: Evidence from aggregated and disaggregated time series data of Pakistan. *Economic Modelling, 31(1), 383-388*. https://doi.org/10.1016/j.econmod.2012.11.035

Jellenz, M., Bobek, V., & Horvat, T. (2020). Impact of education on sustainable economic development in emerging markets—the case of Namibia's tertiary education system and its economy. *Sustainability (Switzerland)*, *12*(21), 1–26. https://doi.org/10.3390/su12218814

Jiburum, U., Nwachukwu, M. U., Mba, H. C., Okonkwo, C. N., & Okeke, D. C. (2021). Determinants of Public Housing Affordability for Large Income Diversity Groups in a New-Town Capital City: A Case Study of Abuja, Nigeria. *SAGE Open*, *11*(3). https://doi.org/10.1177/21582440211045079

Kaasa, Anneli, Factors of Income Inequality and Their Influence Mechanisms: A Theoretical Overview (November 4, 2005). University of Tartu, Faculty of Economics and Business Administration Working Paper No. 40, Available at SSRN: https://ssrn.com/abstract=868491 or http://dx.doi.org/10.2139/ssrn.868491

Ketokivi, M., & McIntosh, C. N. (2017). Addressing the endogeneity dilemma in operations management research: Theoretical, empirical, and pragmatic considerations. *Journal of Operations Management*, *52*, 1–14. https://doi.org/10.1016/j.jom.2017.05.001

Kolstad, I., & Wiig, A. (2015). Education and entrepreneurial success. *Small Business Economics*, *44*(4), 783–796. https://doi.org/10.1007/s11187-014-9621-1

Kumhof, M., Rancière, R., & Winant, P. (2015). Inequality, leverage, and crises. *American Economic Review*, *105*(3), 1217–1245. https://doi.org/10.1257/aer.20110683

International Labour organization (ILO) - Institute of Labour Science and Social Affairs (2014). Improvement of quality, effective utilization and protection of legal rights and interests of wage and salaried workers are the foundation for development of Vietnam’s labour market. Retrieved from https://ilo.org/wcmsp5/groups/public/---asia/---ro-bangkok/---ilo hanoi/documents/publication/wcms_428972.pdf

Li, J., Ding, H., Hu, Y., & Wan, G. (2021). Dealing with dynamic endogeneity in international business research. *Journal of International Business Studies*, *52*(3), 339–362. https://doi.org/10.1057/s41267-020-00398-8

Lu, M., & Yang, W. (2012). Multivariate Logistic Regression Analysis of Complex Survey Data with Application to BRFSS Data. *Journal of Data Science*, *10*, 157–173.

Lv, H. (2017). *The effects of family income on children’s education: An empirical analysis of CHNS data*. *02002*, 49–54. https://doi.org/10.24104/rmhe/2017.04.02002

Marbuah, D. A. (2016). Influence of Parental Income and Educational Attainment on Children ’ s Years of Schooling : Case of Ghana. 1 (Masters thesis, Uppsala Universitet, Sweden). Retrieved from https://uu.diva-portal.org/smash/get/diva2:1092869/FULLTEXT01.pdf

Matthew Semadeni, M. C. W. and S. T. C. (2020). The Perils of Endogeneity and Instrumental Variables in Strategy Research: Understanding through Simulations. Wiley Stable. Strategic Management Journal, 35(7), 1070–1079.https://www.jstor.org/stable/24037334.

Müller, M. (2012). Generalized Linear Models. *Handbook of Computational Statistics* (pp. 681–709). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21551-3_24

National Planning Commission. (2018). Namibia. *Towards Maximising the Demographic Dividend in Namibia*.

Ng'entu Kirima, S. (2019). Effects of Parental Income and Support on Students' Participation in Free Day Secondary Education at Public Day Secondary Schools in Imenti North Sub-County, Kenya. International Journal of Innovative Research and Advanced Studies (IJIRAS), 6(7), 169-173

Ozhamaratli, F., Kitov, O., & Barucca, P. (2022). A generative model for age and income distribution. *EPJ Data Science*, *11*(1). https://doi.org/10.1140/epjds/s13688-022-00317-x

Parker, S. C., & van Praag, C. M. (2006). Schooling, capital constraints, and entrepreneurial performance: The endogenous triangle. *Journal of Business and Economic Statistics*, *24*(4), 416–431. https://doi.org/10.1198/073500106000000215

Reeb, D., Sakakibara, M., & Mahmood, I. P. (2012). From the editors: Endogeneity in international business research. *Journal of International Business Studies*, *43*(3), 211–218. https://doi.org/10.1057/jibs.2011.60

Rehman, S., & Rashidi, M. Z. (2017). The relationship between education and income inequality (the case of Pakistan). *International Journal of Accounting and Economics Studies*, *5*(1), 55. https://doi.org/10.14419/ijaes.v5i1.6505

Royal, T. H. E., Academy, S., & Sciences, O. F. (2022). *Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2022 The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel*. *50005*, 0–72.

Rutz, O. J., & Watson, G. F. (2019). Endogeneity and marketing strategy research: an overview. *Journal of the Academy of Marketing Science*, *47*(3), 479–498. https://doi.org/10.1007/s11747-019-00630-4

Schultz, E. L., Tan, D. T., & Walsh, K. D. (2010). Endogeneity and the corporate governance - performance relation. *Australian Journal of Management*, *35*(2), 145–163. https://doi.org/10.1177/0312896210370079

Semadeni, M., Withers, M. C., & Trevis Certo, S. (2014). The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations. *Strategic Management Journal*, *35*(7), 1070–1079. https://doi.org/10.1002/smj.2136

Sivakumar, M.A & Sarvalingam, M.A. (2015). Human Deprivation Index: A Measure of Multidimensional Poverty. Munich Personal RePEc Archive, 22337(61160), 1–53. Retrieved from https://mpra.ub.uni-muenchen.de/22337/1/MPRA_paper_22337.pdf

Stock, J. H., & Watson, M. W. (2019). Introduction to Econometrics (4th Edition). New York, United States of Ameria:Pearson Addison Wesley

Stryzhak, O. (2020). The relationship between education, income, economic freedom and happiness. *SHS Web of Conferences*, *75*, 03004. https://doi.org/10.1051/shsconf/20207503004

Terza, J. V. (2016). *Two-Stage Residual Inclusion Estimation: A Practitioners Guide to Stata Implementation*. *46202*.

Teweldemedhin, M. (2015). Factors Influencing Income Inequality in Namibia. *British Journal of Economics, Management & Trade*, *10*(4), 1–14. https://doi.org/10.9734/bjemt/2015/19249

The World Bank Annual Report 2020. (2020). In *the World Bank Annual Report 2020*. https://doi.org/10.1596/978-1-4648-1619-2

Udokang Anietie, M. A., Raji Surajudeen, E., & Bello Latifat Kemi, T. (2015). An Empirical Study of Generalized Linear Model for Count Data. *Journal of Applied & Computational Mathematics*, *04*(05). https://doi.org/10.4172/2168-9679.1000253

Ullah, S., Akhtar, P., & Zaefarian, G. (2018). Dealing with endogeneity bias: The generalized method of moments (GMM) for panel data. *Industrial Marketing Management*, *71*, 69–78. https://doi.org/10.1016/j.indmarman.2017.11.010

Vo, D., Nguyen, T., Tran, N., & Vo, A. (2019). What Factors Affect Income Inequality and Economic Growth in Middle-Income Countries? *Journal of Risk and Financial Management*, *12*(1), 40. https://doi.org/10.3390/jrfm12010040

Wintoki, M. B., Linck, J. S., & Netter, J. M. (2012). Endogeneity and the dynamics of internal corporate governance. *Journal of Financial Economics*, *105*(3), 581–606. https://doi.org/10.1016/j.jfineco.2012.03.005

Wolla, S. A., & Sullivan, J. (2017). Education, Income, and Wealth. *Federal Reserve Bank of St. Louis Research*, *January*. https://doi.org/10.5860/choice.51-1007

Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach* (5th ed.). United States of America, South-Western, Cengage Learning

Zaefarian, G., Kadile, V., Henneberg, S. C., & Leischnig, A. (2017). Endogeneity bias in marketing research: Problem, causes and remedies. *Industrial Marketing Management*, *65*, 39–46. https://doi.org/10.1016/j.indmarman.2017.05.006

Zanin, L., Radice, R., & Marra, G. (2015). Modelling the impact of women's education on fertility in Malawi. *Journal of Population Economics*, *28*(1), 89–111. https://doi.org/10.1007/s00148-013-0502-8

Zare, R. (2019). Globalization and Income Inequality: Does the Level of Financial Development Matter ? 1 Introduction. *Journal of Money and Economy*, *14*(2), 159–175.

APPENDICES

APPENDIX 1. Stata Codes: Data Management (Validation, Cleaning and Processing) and Estimation of Income

//This Do File analyses the 2015/2016 Namibia Housing Income and Expenditure Survey using the file titled “ind_level_tabcomb_salary18to60yrs”//

*This file has codes for all the stages of the analysis, outlined here below:

1. DEFINE WORKING DIRECTORIES

2. DECLARING THE SURVEY DESIGN

3. INSTALLING PACKAGES

4. GENERATING VARIABLES

5. REGRESSION ANALYSIS: ORDINARY LEAST SQUARE EQUATION

6. REGRESSION ANALYSIS: TWO STAGE-LEAST REGRESSION

7. PERFORMING TESTS OF ENDOGENEITY

****

//Declaring the survey design

* Uploading the original dataset

use "$input", clear

use "C:\Users\mnanyemba\Desktop\11MSSB UNAM 2019\MSC THESIS WRITE UP 2021\NHIES Data\ind_level_tabcomb_salary18to60yrs.dta"

*Survey Setting

svyset psu_id [pweight=wgt_hh], strata(stratum) vce(linearized) singleunit(missing)///svyset for the household

svyset psu_id [pweight=wgt_ind], strata(stratum) vce(linearized) singleunit(missing)///svyset for the individual

svyset psuno_f [pweight = wgt_ind],strata(stratum)

gen q10_02_1_log=log(q10_02_1)

******************Installing packages************

ssc install ivhettest ///module to perform Pagan-Hall and related heteroskedasticity tests after IV

ssc install ivendog ///module to calculate Durbin-Wu-Hausman endogeneity test after ivreg

ssc install ivrobust

ssc install ivolsdec

ssc install ivregress2

ssc install ivvif

**Generating some variables by combining responses of certain variables.

******************dwelling unit type*************************

gen dwelling_unit_type =.

replace dwelling_unit_type = 1 if q02_01==1

```
replace dwelling_unit_type = 2 if q02_01==8

replace dwelling_unit_type = 3 if q02_01==9

replace dwelling_unit_type = 4 if q02_01==2 |q02_01==3|q02_01==4

replace dwelling_unit_type = 5 if inlist(q02_01,5,6,7,10) & q02_01!=.

///means type of dwelling should not be equal to missing.

#delimit

lab def dwelling_unit_type

1 "Detached house"

2 "Traditional dwelling"

3 "Improvised housing unit"

4 "Town house/Apartment"

5 "Others" ;

#delimit cr

lab value dwelling_unit_type dwelling_unit_type

///this var represent variable

label variable dwelling_unit_type "Dwelling unit type"


*****************main material used for the wall*****************

gen main_material_wall =.

replace main_material_wall = 1 if q02_03==1
```

```
replace main_material_wall = 2 if q02_03==3

replace main_material_wall = 3 if q02_03==4 |q02_03==5

replace main_material_wall = 4 if inlist(q02_03,2,6,7,8,9,10,11) & q02_03!=.

///means main_material_wall should not be equal to missing.

#delimit

lab def main_material_wall

1 "Cement blocks/Bricks/Stones"

2 "Corrugated iron/Zinc"

3 "Sticks, wooden poles, clay and/or cow-dung"

4 "Others" ;

#delimit cr

lab value main_material_wall main_material_wall

///this var represent variable

label variable main_material_wall "Main material wall"


************main source of energy*********************************

gen main_source_energy_lighting =.

replace main_source_energy_lighting = 1 if q02_09==1

replace main_source_energy_lighting = 2 if q02_09==13

replace main_source_energy_lighting = 3 if q02_09==7
```

```
replace main_source_energy_lighting = 4 if q02_09==9

replace main_source_energy_lighting = 5 if inlist(q02_09,2,3,4,5,10,11) & q02_09!=.

///means type of dwelling should not be equal to missing.

#delimit

lab def main_source_energy_lighting

1 "Electricity from mains"

2 "Battery lamp/ Torch/ Cell phone"

3 "Candles"

4 "Solar energy"

5 "Others" ;

#delimit cr

lab value main_source_energy_lighting main_source_energy_lighting

///this var represent variable

label variable main_source_energy_lighting "Main source of energy for lighting"


************************internet access******************************

gen internet_access =.

replace internet_access = 1 if inlist(q05_02_38,1,2,3)

replace internet_access = 2 if q05_02_38==4 & q05_02_38!=.

///means access to internet should not be equal to missing.
```

```
#delimit

lab def internet_access

1 "Internet access"

2 "No internet access";

#delimit cr

lab value internet_access internet_access

///this var represent variable

label variable internet_access " Household Internet access"



**********************marital status **************************

gen marital_status =.

replace marital_status = 1 if q01_07==1

replace marital_status = 2 if inlist(q01_07,2,3,4)

replace marital_status = 3 if inlist(q01_07,5,6,7) & q01_07!=.

///means marital status should not be equal to missing.

#delimit

lab def marital_status

1 "Never married"

2 "Married"

3 "Others" ;
```

```
#delimit cr

lab value marital_status marital_status

///this var represent variable

label variable marital_status "Marital status"



*****************Employer type***********************************

gen employer_type =.

replace employer_type = 1 if q06_29_1==1 | q06_29_1==3

replace employer_type = 2 if q06_29_1==2

replace employer_type = 3 if q06_29_1==4

replace employer_type = 4 if q06_29_1==5 | q06_29_1==6 & q06_29_1!=.

///means employer_type should not be equal to missing.

#delimit

lab def employer_type

1 "Private company, State-owned enterprise, commercial"

2 "Government, public sector or army "

3 "Private household/ individual"

4 "Others" ;

#delimit cr

lab value employer_type employer_type
```

```
///this var represent variable

label variable employer_type "Employer type"


********************work type****************************

gen work_type =.

replace work_type = 1 if q06_19_1==1

replace work_type = 2 if q06_19_1==2

replace work_type = 3 if inlist(q06_19_1,3,4,5,6) & q06_19_1!=.

///means work_type should not be equal to missing.

#delimit

lab def work_type

1 "Wage/ Salary Activity: In agriculture"

2 "Wage/ Salary Activity: Not in agriculture"

3 "Others" ;

#delimit cr

lab value work_type work_type

///this var represent variable

label variable work_type "Work type"


********************employment status ******************
```

```
gen employment_status =.

replace employment_status = 1 if employment_status_broad==1

replace employment_status = 2 if employment_status_broad==2 | employment_status_broad==3 & employment_status_broad!=.

///means employment status should not be equal to missing.

#delimit

lab def employment_status

1 "Employed"

2 "Not employed" ;

#delimit cr

lab value employment_status employment_status

///this var represent variable

label variable employment_status "Employment status"


*********************employ employed**********************

gen employ_period =.

replace employ_period = 1 if q06_35_1==1

replace employ_period = 2 if q06_35_1==2 | q06_35_1==3

replace employ_period = 3 if q06_35_1==4

replace employ_period = 4 if q06_35_1==5 & q06_35_1!=.
```

```
///means employment period should not be equal to missing.

#delimit

lab def employ_period

1 "Less than 1 year "

2 "1-5 years "

3 "6-10 years "

4 "6 years or more" ;

#delimit cr

lab value employ_period employ_period

///this var represent variable

label variable employ_period "Employ period"



************************total hours****************************

gen totalhrs = .

replace totalhrs  = 1 if inrange(q06_22_tot_1,1,8) & q06_22_tot_1!=. //total hours worked is not equal to missing

replace totalhrs  = 2 if q06_22_tot_1>8 & q06_22_tot_1!=.

#delimit ;

lab def totalhrs

1 "1-8"
```

```
2 "8+" ;

#delimit cr

label val totalhrs totalhrs //this val represent value

lab var totalhrs "totalhrs" //this var represent variable



************Mode of transport*************

gen transport_mode =.

replace transport_mode = 1 if q06_39_1==1

replace transport_mode = 2 if q06_39_1==6 | q06_39_1==7

replace transport_mode = 3 if q06_39_1==4 | q06_39_1==5

replace transport_mode = 4 if inlist(q06_39_1,2,3,8,9,10,11) & q06_39_1!=.

///means transport_mode should not be equal to missing.

#delimit

lab def transport_mode

1 "Walk "

2 "Public (taxi/minibus/bus) "

3 "Private (sedan/bakkie) "

4 "Others" ;

#delimit cr

lab value transport_mode transport_mode
```

```
///this var represent variable

label variable transport_mode "Transport_mode"

*************Running weighted data - categorical variables****************8

tab q01_02 [iw=wgt_hh],m    ///to get the number of households in the analysis

tab q01_02 [iw=wgt_ind],m   ///to get the number of individuals in the analysis

tab main_sourceincome [iw=wgt_hh],m

 tab  q02_01 [iw=wgt_hh],m

tab  q02_01 [iw=wgt_ind]


*********************tabulating the categorical variables************

tab region [iw=wgt_ind]

tab urbrur [iw=wgt_ind]

tab dwelling_unit_type[iw=wgt_hh]

tab main_material_wall[iw=wgt_hh]

tab main_source_energy_lighting[iw=wgt_hh]

tab internet_access [iw=wgt_hh]

tab q20_04[iw=wgt_hh]

tab q01_02 [iw=wgt_ind]

tab marital_status [iw=wgt_ind]

tab employer_type [iw=wgt_ind]
```

tab work_type [iw=wgt_ind]

tab q03_01[iw=wgt_hh]

tab q03_09_mj[iw=wgt_hh]

***************tabulating mean income with other variables**********

svy: mean q10_02_1, over(region)

svy: mean q10_02_1, over(attain)

svy: mean q10_02_1, over(totalhrs)

svy: mean q10_02_1, over(main_material_wall)

svy: mean q10_02_1, over(employer_type)

svy: mean q10_02_1, over(employ_period)

svy: mean q10_02_1, over(work_type)

svy: mean q10_02_1, over(urbrur)

svy: mean q10_02_1, over(q03_04_mj)

svy: mean q10_02_1, over(q01_02)

svy: tab main_sourceincome over(region)

******************REGRESSION ANALYSIS******************

***Running the regression model: Basic Ordinary Least Square - without considering the clustering, stratification and weighting regress q10_02_1_log i.region i.dwelling_unit_type i.main_material_wall i.q20_04 q01_06_y i.q01_02 hh_members i.marital_status i.employer_type i.work_type q03_05 i.internet_access i.attain i.main_source_energy_lighting q02_51_05 i.totalhrs i.employ_period i.transport_mode

Running the regression model considering the clustering, stratification and weighting

svy linearized : regress q10_02_1_log i.region i.dwelling_unit_type i.main_material_wall i.q20_04 q01_06_y i.q01_02 hh_members i.marital_status i.employer_type i.work_type q03_05 i.internet_access i.attain i.main_source_energy_lighting q02_51_05 i.totalhrs i.employ_period i.transport_mode

********TESTING FOR ENDOGENEITY********

****************Running the Two-Stage Least Regression***************

svy linearized : ivregress 2sls q10_02_1_log i.region i.dwelling_unit_type i.main_material_wall i.q20_04 q01_06_y i.q01_02 hh_members i.marital_status i.employer_type i.work_type i.totalhrs i.employ_period i.transport_mode (q03_05 = i.internet_access i.attain i.main_source_energy_lighting q02_51_05)

ivregress 2sls q10_02_1_log i.region i.dwelling_unit_type i.main_material_wall i.q20_04 q01_06_y i.q01_02 hh_members i.marital_status i.employer_type i.work_type i.totalhrs i.employ_period i.transport_mode (q03_05 = i.internet_access i.attain i.main_source_energy_lighting q02_51_05) test

estat endogenous

estat firststage

estat overid

hettest, rhs fstat

*******************GRAPHICS********************************

graph matrix q10_02_1_log q03_05 hh_members q01_06_y q06_22_tot_1 ///producing a scatterplot matrix

graph matrix q10_02_1_log q03_05, title (Scatter plot of Income against Education)

graph scatter q10_02_1_log q03_05

scatter q10_02_1_log q03_05 hh_members q01_06_y q06_22_tot_1

scatter q10_02_1_log q03_05 || lfit q10_02_1_log q03_05 ///producing a scatterplot for income and education

scatter q10_02_1 q03_05 || lfit q10_02_1 q03_05

//////producing a scatterplot for income (before log) and education

plot q10_02_1_log q03_05

//Drawing a bar graph of main source of income

graph bar, over(main_sourceincome)

APPENDIX II. The OLS and 2SLS model coefficients/estimates

Table 17: Comparison of OLS and 2SLS model coefficients

Variable	**OLS**		**2SLS**	
Region	**Coef.**	**Std. Err.**	**Coef.**	**Std. Err.**
Erongo	-0.2801	0.0731	-0.5138	0.1444
Hardap	-0.1109	0.0785	-0.1754	0.1290
Kavango East	-0.8457	0.1061	-0.8894	0.2717
Kavango West	-0.6461	0.1116	-0.7111	0.1714
Khomas	-0.1088	0.0691	-0.0192	0.0975
Kunene	-0.1208	0.1115	-0.2346	0.1155
Ohangwena	-0.9090	0.1037	-0.9772	0.1633
Omaheke	-0.5885	0.0934	-0.7164	0.2205
Omusati	-0.5827	0.1076	-0.7479	0.1763
Oshana	-0.5175	0.0837	-0.7501	0.1524
Oshikoto	-0.1552	0.0918	-0.1421	0.1416
Otjozondjupa	-0.3107	0.0737	-0.2433	0.1063
Zambezi	-1.4472	0.1068	-1.6696	0.2174
//Kharas (reference group)				
Dwelling unit type	-0.1266	0.0797	-0.3566	0.1388
Traditional dwelling	0.1084	0.0635	0.0926	0.1263
Improvised housing unit	-0.0123	0.0514	-0.1274	0.0932
Town house/Apartment	-0.2364	0.0813	-0.6461	0.2784
Detached house (reference group)				
Main material for the wall				
Corrugated iron/Zinc	-0.3207	0.0599	-0.5697	0.1326
Sticks, wooden poles, clay and/or cow dung	-0.2574	0.0785	-0.2879	0.1506
Others	0.0757	0.0708	0.1909	0.1147
Cement blocks/Bricks/Stones (reference group)				
Contribution to investment				
No	-0.1597	0.0357	-0.1079	0.0635
Yes (reference group)				
Age of respondent	0.0091	0.0020	0.0125	0.0037
Sex of respondent				
Male	0.3885	0.0489	0.4211	0.0516

Female (reference group)				
Household size	-0.0405	0.0158	-0.0382	0.0163
Marital status				
Married	0.4339	0.0593	0.4071	0.0637
Others	0.3416	0.1103	0.3435	0.1085
Never married (reference group)				
Type of employer				
Government, public sector or army	0.7857	0.0757	0.7125	0.0899
Private household/ individual	-0.6786	0.0894	-0.6055	0.0943
Others	-0.0169	0.1986	-0.0704	0.2042
Private Company, State-owned enterprise, Commercial (reference group)				
Type of work		0.0901		
Wage/ Salary Activity: Not in agriculture	0.0790	0.0090	0.3790	0.1060
Wage/Salary Activity: In Agriculture (reference group)				
Education	0.0468	0.0056	0.1270	0.0243
Total hours worked				
8+	0.8109	0.1416	0.8352	0.3768
Less than 8 working hours (reference group)				
Employment period				
1-5 years	0.9077	0.0418	0.9767	0.0716
6-10 years	1.0440	0.0584	1.1536	0.0949
11 years or more	1.2268	0.0613	1.2864	0.1042
Less than 1 year (reference group)				
Transport mode				
Public (taxi/minibus/bus)	0.3659	0.0470	0.3338	0.0772
Private (sedan/bakkie)	0.5359	0.0476	0.5411	0.0861
Others	0.2344	0.0930	0.2263	0.1235
Walk (reference group)				
_cons	7.2741	0.2023	6.5162	0.4927

APPENDIX III. Regional boundary map of Namibia