THE EFFECTS OF WEIGHTING IN THE REGRESSION ANALYSIS OF SURVEY DATA COLLECTED USING NON-PROBABILISTIC SAMPLING METHODS: A SECONDARY DATA ANALYSIS

UNIVERSITY OF ZIMBABWE

COLLEGE OF HEALTH SCIENCES

DEPARTMENT OF COMMUNITY MEDICINE

BY

LEON-SAY MUDADI

(R141480W)

SUPERVISOR: PROFESSOR S. RUSAKANIKO

MASTERS IN BIOSTATISTICS (COMMUNITY MEDICINE)

August 2015

DECLARATION FORM
STUDENT

I do hereby declare that this dissertation is the original work of LEON-SAY MUDADI and has not been submitted before to the University of Zimbabwe or any other institution for the fulfilment of any degree requirements.

Name.....**Leon-Say Mudadi**.....

Signed.....Date.....

SUPERVISOR

I certify that I have supervised the writing of this dissertation and declare that it is indeed the original work of the student in whose name it is being submitted.

Name.....**Professor S. Rusakaniko**.....

Signed..... Date.....

DEPARTMENTAL CHAIRPERSON

I do hereby declare all the above statements to be true.

Name.....**Professor S. Rusakaniko**.....

Signature.....Date.....

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Prof S Rusakaniko for his valuable academic guidance throughout this project. His advice and critical comments have been of great value to me. My sincere gratitude goes to my Biostatistics lecturers: Dr W Tinago, Dr G Mandozana and Mr V Chikwasha for the mentorship they rendered to me to complete my thesis.

A special thank you goes to my fellow classmates who assisted me throughout this project. Last but not least, I would like to express my sincere gratitude to my wife and son for their love and encouragement.

ABSTRACT

Introduction

When surveys are conducted especially for hidden populations, data is rarely collected using random sampling which is the ideal way to collect representative data. However, it is common practice to analyse this data as if it was collected through random sampling ignoring the sampling design. We sought to determine the effects of including weights in the analysis of survey data collected through non-probabilistic sampling methods.

Broad objective

To assess the effects of weighting on risk taking behaviours associated with STIs among female sex workers (FSW) and long distance truck drivers (LDTD) in Beitbridge using weighted and unweighted logistic regression models.

Methods

Both inverse probability weighted and unweighted forward selection multivariate logistic modelling techniques were used to determine significant risk taking behaviours associated with STIs in FSW and LDTD. Final models compared magnitude of the difference between odds ratios, the selection of final variables, standard errors, statistical significance of selected variables and the overall fit of the models to determine whether or not we believed weighted models were more appropriate for the analysis of the survey data for FSW and LDTD.

Results

For risk taking behaviours associated with STIs, inclusion of weights resulted in an increase in the odds ratios, a decrease in the standard errors and narrowing of the confidence intervals for the parameters in the weighted model for FSW. In the weighted model for LDTD, the odds ratios were higher than in the unweighted model and the confidence intervals were slightly narrow. However, the standard errors were higher in the weighted models.

Conclusion

Based on the results, we concluded that weighting in the regression analysis of survey data collected using non probabilistic sampling methods helps to improve the precision of the regression estimates; hence weighted models should be used.

Table of Contents

LIST OF TABLES AND FIGURES

LIST OF ABBREVIATIONS

AIDS Acquired Immune Deficiency Syndrome

ART Anti Retroviral Treatment

CDC Centre for Disease Control and prevention

HIV Human Immunodeficiency Virus

IDU Intravenous Drug Use

LDTD Long Distance Truck Drivers

MARPs Most At Risk Populations

PMTCT Prevention of Mother To Child Transmission

RDS Respondent Driven Sampling

SADC Southern African Development Community

STI Sexually Transmitted Infections

FSW Female Sex Workers

TLS Time Location Sampling

WHO World Health Organisation

CHAPTER ONE

1. INTRODUCTION

1.1. Sampling hard to reach populations

Non probabilistic sampling methods have become desirable in studies focusing on hard to reach (hidden) populations such as illegal drug users, sex workers and men who have sex with men in view of the HIV/AIDS pandemic. This is due to the fact that hard to reach population subgroups play an important role in the transmission of the HIV virus in human populations, but it is often difficult to investigate the factors that drive transmission in these groups by using representative (probability) samples[1]. Absence of an adequate sampling frame coupled with the association between these hard to reach groups and illicit activities or stigma favours the use of non probabilistic sampling methods to investigate factors that drive transmission of sexually transmitted infections (STIs) and HIV[1].

Just like female sex workers (FSW), long distance truck drivers (LDTD) also fall in the category of hard to reach populations. In order to undertake studies on these groups of hard to reach populations, non-probabilistic/quasi sampling methods such as snowball sampling, respondent driven sampling (RDS) and time location sampling (TLS) are used[2]. These sampling methods were devised since adequate surveillance of hard to reach subpopulations is crucial to containing the HIV epidemic in low prevalence settings and in slowing the rate of transmission in high prevalence settings[3]. It is important to understand the risk taking behaviours of these hidden populations that drive the transmission of HIV and other STIs.

LDTD and their commercial sex contacts (women and men with whom they exchange money and/drugs for sex) have been implicated in the spread of HIV and other STIs along major transportation routes in developing countries[5]. However, there is a paucity of literature on the HIV/STI risk taking behaviours of LDTD and FSW.

Several reasons based on studies done in Eastern Africa on why LDTD and FSW maybe at more increased risk of HIV/STI infection have been postulated. From the studies, it was found out that LDTD normally spend nights away from home and because of this they may engage in sex with many casual and semi-regular

partners along the transport routes they travel, sometimes without using condoms[5]. Results from studies of LDTD suggest that they have low HIV/STI knowledge, have higher reported rates of STIs, engage in sex with multiple regular and commercial partners while on the road, report low condom use and engage in illicit drug use[6]. These behaviours put them at a greater risk of acquiring or transmitting HIV/STIs[9]. Likewise, FSW who live or operate along major transport routes engage in sex with multiple partners, usually in situations of non or inconsistent condom use[5]. The need for survival usually drives FSW into unprotected sex with many men, thereby exposing themselves to the risk of acquiring or transmitting HIV/STIs in the process[10].

Intravenous drug use (IDU), alcohol abuse and unprotected sexual intercourse with regular non paying clients have been shown to be some of the high risk behaviours that FSW are involved in[10]. As a result of these behaviours, FSW are at an increased risk of contracting HIV due to multiple exposures: large number and concurrency of sexual partners, inconsistent condom use, intersection with injecting drug use and presence of other STIs[9]. However, in Southern African settings intravenous use of drugs such as heroin and other stimulants as well as addiction is still very low[7]. Nevertheless, for those FSW addicted to drugs, sex work becomes their only source of income and they will do anything a client demands just to get money to feed the addiction, or they trade sex for drugs[4]. More often, this implies having unprotected sex with a client thereby increasing their risk of contracting STIs and HIV.

When surveys are conducted for hard to reach populations using non-probabilistic/quasi sampling methods, it is desirable to make results of such surveys representative of the target population[3]. This is usually achieved by considering the complex sampling design during data analysis which will accord the appropriate weights so as to produce unbiased estimates, or more realistic estimates with minimal level of bias[8]. However, absence of a full sampling frame as well as inability to have full control of the sampling procedure by the sampler always introduces a bias which will need to be corrected by sample calibration/balancing to re-weight sample data so that it becomes representative.

1.2. Primary data

In 2013, the World Health Organisation (WHO) conducted an HIV Sero-Behavioural survey in the Southern African Development Community (SADC) region. The survey was targeted at FSW and LDTD in border towns. Data was collected on a range of variables that include demographics, sexual practices, alcohol and drug use, signs and symptoms of STIs as well as knowledge of HIV transmission methods. This data was captured into STATA®.

1.3. Background to the original research

1.3.1. Introduction

Behavioural and serological surveys have increasingly become common in border areas due to their enabling environment for HIV transmission. As most of these border areas are rural and far from urban centres, access to socio-economic services such as formal employment and education is usually limited. These have often attracted risk practices such as commercial sex. Despite the known risk factors in border areas, data on the magnitude of HIV and STI among FSW and LDTD in the border areas in SADC region is still limited.

1.3.2. Background of the study

SADC region has the highest burden of HIV and largest number of people in need of HIV treatment and care in the world. Most of its countries have generalized epidemics though some like Madagascar, Seychelles and Mauritius have concentrated epidemics driven by key populations also known as Most At Risk Populations (MARPs). In SADC, increased cross border movements are largely due to variations in economic developments among member states and land locked nature of some of its members. LDTD, FSW and mobile/cross border traders have been common on this route. Due to the transient nature of these populations, cross border areas often provide a good opportunity for risky sexual behaviours such as transactional and commercial sex for mobile and resident populations.

1.3.3. Problem statement of the original study

Border areas have long been identified as environments of high HIV vulnerability. The association of mobility, migration and engagement of high risky sexual behaviour and high HIV prevalence at border areas has been observed in other countries in sub-Saharan Africa, Asia and South America. However, data on the burden of HIV in SADC countries border areas is limited, but it is highly likely that the prevalence of HIV and STI will be high as well due to the known risk factors in border areas. The magnitude of HIV and STIs among FSW and LDTD in the border areas in SADC region remain largely unknown thus warranting investigation.

1.3.4. Research question of the original study

What is the magnitude of HIV and STIs among FSW and LDTD in 3 border sites covered in the survey?

1.3.5. Objectives of the original study

Primary objective of the original study

To measure the prevalence of HIV and STIs and their related risk behaviours among FSW and LDTD in 3 border sites of Beitbridge, Ngwenya and Wenela in the SADC region.

Specific objectives of the original study

- To asses sexual and other risk behaviours associated with HIV and STI transmission among FSW and LDTD in 3 border sites
- To describe demographic characteristics related to HIV and STI infections in FSW and LDTD
- To determine the proportion of FSW and LDTD in 3 border sites who were infected by Syphilis (Syphilis as a biological marker of STI infections)
- To assess the coverage of selected HIV and STI services (HIV testing and counselling, STI treatment, PMTCT and ART/care) among FSW and LDTD in 3 border sites.

1.3.6. Study design of the original study

The study design was a cross sectional epidemiological survey with behavioural and biological markers.

1.3.7. Sample size of the original study

The sample size at each border site was calculated by factoring in a behavioural and biological component, design effect of 2, significance level of 0.05 and a power of 80%. A sample size of 330 FSW at each border site was calculated. For Swaziland, a sample size of 355 FSW at each border site was calculated. For LDTD, a sample size of 460 at each border site was calculated. The sample sizes at each border site were adjusted upwards to account for a non-response rate of 10%.

1.3.8. Sampling methods of the original study

Border sites

Border sites that were included in the study were selected purposively basing on the following criteria:

- High number of cross-border activities such as transportation and trade
- Sites where project clinics operate
- High burden of HIV from previous assessments, reports or research

The border sites that were included in the final study are Beitbridge border post (Zimbabwe), Ngwenya border post (Swaziland) and Wenela border post (Namibia).

Female sex workers

Respondent driven sampling (RDS) was used to recruit FSW at each border site. RDS is a quasi-probability sampling method based around a structured form of snowball sampling. Unlike snowball sampling, the use of RDS enables the sample to resemble the population of interest. It involves first recruiting key respondents (‘seeds’) with large networks who are then asked to recruit a further three new recruits. These new recruits are also each asked to recruit a further three new recruits, and so on, until the desired sample size is reached.

Recruitment is facilitated by a monetary reward system in which each person receives a sum of money for their participation as a reimbursement for their travel costs. For this survey, each participant received some phone credit (airtime) for each eligible peer that they recruited. This credit was to reimburse participants for recruiting their peers. RDS is an effective way of recruiting 'hidden' populations provided those populations are relatively well networked.

Long distance truck drivers

Time location sampling (TLS) was used to recruit LDTD. It is a form of cluster sampling with both location and time components such that eligible populations who access specific locations at a certain point in time will be reached for the survey. The sampling began with identification and listing of time location sites at the border areas for truck drivers (formal or informal truck parks along the road/highway including peak hours). As most border sites are likely to have small geographical area with small populations, all identified time location sites were selected for the survey. All eligible individuals in the selected time location sites were then recruited for interviews and collection of blood samples until the sample size was reached.

1.3.9. Data collection methods of the original study

The data collection process commenced with the local recruitment of data collectors who were mainly health workers such as nurses, clinical assistants, laboratory workers or nursing assistants. All members of the survey team went through extensive training in conducting RDS and TLS surveys as well as using Personal Digital Assistants (PDAs). Questionnaire was pre-tested to ensure consistency and onsite trouble shooting related to PDAs was done as well. Survey participants were asked to provide written consent to be interviewed and to have blood samples drawn.

The questionnaire collected information on demographics, knowledge, condom use and risk behaviours related to HIV transmission as well as cross-border movements and access to basic HIV and STI services, HIV testing and counselling and STI treatment. Blood samples collected were to be tested for HIV and Syphilis.

Data was collected electronically using PDAs and sent directly to the within country central point. This data was part of a multicentre study where the sample size at each site had enough power to allow country specific analysis.

1.3.10. Data management and analysis techniques of the original study

Data from PDAs was uploaded directly to the Microsoft Access database. The data had been checked for completeness and consistency before being uploaded to the database. From the database, the data was then exported to statistical software for advanced statistical analysis. A back up database was generated by Site coordinators in collaboration with the survey statistician.

According to the survey protocol, Respondent Driven Sampling Analysis Tool (RDSAT) was to be used to apply the appropriate weighting to adjust for biases found in RDS chain referral sample. During analysis, RDSAT accounts for the snowball-like initial selection of seeds as well as network influence of the seeds and uses a weighting system in order for an RDS obtained sample to be considered probabilistic and representative of the social network the sample was recruited from. Analysis techniques for TLS were not mentioned.

1.4. Critical appraisal of the original study

1.4.1. Problem statement

The problem statement is quite clear; the magnitude of HIV and STI prevalence amongst LDTD and FSW is not known but is assumed to be higher than in the general population. A recent study found out that the prevalence of HIV among FSW was between 50% and 70%[10], but there is a possibility that it could be higher amongst those based in border areas since there is a documented link between mobility and HIV vulnerability[9]. No study has quantified the prevalence of HIV and STIs amongst LDTD, despite them being most at risk of contracting and transmitting HIV and STIs due to the nature of their job and tendency to indulge in risky sexual behaviours. In addition, the objectives are in line with the problem statement.

1.4.2. Objectives

The objectives are specific, measurable, achievable, realistic and time bound (**SMART)**:

Specific – The objectives clearly state what was to be achieved by this survey. The primary objective was to measure the prevalence of HIV and STIs and their related risk behaviours among FSW and LDTD in three border sites. In addition, the objective explicitly states the target group, in this case FSW and LDTD.

Measurable – The objectives were set to determine the baseline prevalence of HIV among FSW and LDTD in the 3 border sites. These findings were to be used to assess trends and changes in the prevalence of HIV in future surveys.

Achievable – Objectives were achievable since adequate resources were made available by the World Health Organisation.

Realistic – During the course of the survey, the prevalence of HIV among FSW and LDTD was going to be determined, hence making the objectives realistic. This was in keeping with the scope of the survey.

Time bound – All the objectives were to be achieved within the timeframe of the survey as was set out in the work plan, hence they were time bound.

1.4.3. Study design

Since the primary objective of the study was to determine the prevalence of HIV and STIs amongst the target group, the use of the cross sectional design was appropriate.

1.4.4. Sample size

The sample size for both FSW and LDTD was adequately calculated. The power was set at 80% for both samples. In addition, the factoring in of behavioural as well as biological component in calculation of the sample size ensured a more powered representative sample.

1.4.5. Sampling

Sex workers are hard to reach populations and hence recruiting them into studies using the usual chain referral methods introduces a bias that is related to the chain of influence that a single sex worker has. To avoid this bias, RDS was used for this study and one sex worker was allowed to recruit only 3 other sex workers.

Researchers would keep track of who recruited whom to avoid over-sampling of well connected individuals. This makes it possible to draw statistically valid samples of this hard to reach group.

Time location sampling (TLS) was used to recruit LDTD into the study. This sampling technique was adequate because there is no comprehensive list of LDTD and they happen to be at different sites at different times. Thus as a result of using TLS each LDTD has a non-zero probability of being included into the study.

1.4.6. Data collection methods

A questionnaire was used to collect data and this was coupled with drawing of blood samples for HIV and STI screening. For the purposes of the survey, the data collection methods were appropriate as they would give enough information about the target population. However, some questions which are critical during data collection for respondent driven samples were not included in the questionnaire. This has an effect on the analysis methods that will have to be used.

1.4.7. Data analysis techniques

Data was to be analyzed using the RDSAT software for analyzing data collected using RDS technique for FSW. However, the data does not meet the assumptions that enable the use of RDSAT for analysis. The first assumption is of reciprocity, which implies that if respondent A recruited respondent B, then in principle, B could have recruited A[7]. In practice this assumption is tested by including a survey question about the relationship between the respondent and the recruiter. The assumption is violated if many of the recruited persons are strangers. Within the survey questionnaire, the question to test the assumption was not included; hence this assumption was not met thereby making the RDSAT analysis inappropriate.

The second assumption is to do with networked population like that of sex workers. It is assumed that all the respondents are interconnected. This assumption would be violated for example if the target population consisted of rivalling gangs who do not communicate with one another, a common phenomenon in commercial sex work. In such a scenario, separate RDS samples

for each of the non communicating groups have to be conducted. This was not considered in the survey and hence this assumption was not met.

The third assumption is degree, where respondents accurately report their degree in the network, this assumption was again not met since this question was not asked during data collection.

1.5. Problem statement

LDTD and FSW are hard to reach populations. Sampling of these populations does not usually follow simple random sampling. Time location sampling was used to sample LDTD while respondent driven sampling was used to sample FSW. Analysis of data collected through these sampling techniques is often complex, since it has to factor in the sampling design to accord **appropriate weights** to reduce bias.

An important aspect of RDS sampling is to determine the size of the social network of the seeds. Once this is known, the Volz – Heckathorn estimator (RDS-2) will be used for weighting data collected through RDS to account for the sampling design during analysis. Since it has been shown that the degree of the 'seed' was not determined during the data collection process for female sex workers, it means the data could not be analysed using RDSAT.

To analyze data collected through TLS, there is need to factor in weights which correspond to the frequency of an individual at a venue so as to reduce bias especially in the event where there might be a positive association between frequency of attendance and an outcome of interest. This component was not considered during data collection hence analyzing data for LDTD using frequency weighting for TLS data is not possible.

Despite the aforementioned pitfalls in the data, analysis of this data using other statistical analysis techniques for survey data provides an opportunity to understand the risk taking behaviours of LDTD and FSW in Beitbridge which are associated with contracting STIs. Weighted logistic regression for survey data has been shown to be an alternative analysis technique when dealing with complex survey data[11] hence it will be used to analyze data for LDTD and FSW.

CHAPTER TWO

2. Literature review

2.1. Risk taking behaviours of LDTD and FSW

Although recent trends in Zimbabwe suggest a declining prevalence of HIV, reports indicate that as many as 40 000 Zimbabweans contract the virus each year[12]. It is widely believed that core groups - that is, groups of individuals who have large numbers of sex partners who themselves have large numbers of sex partners play an important role in the spread and persistence of HIV and STIs and are characterized by a high prevalence of STIs and HIV[13]. FSW and LDTD are examples of such core groups. As a result of their risky sexual practices, these core groups also act as bridge populations that is individuals who have sexual links with members of both high and low STI/HIV prevalence subpopulations- thus playing a big role in transmitting infections from the core groups to the general population[14].

According to the Centre for Disease Control and Prevention (CDC), there has been a few population based studies that have been done on HIV risk and sex workers[15]. However, it is known that the risk of HIV and other sexually transmitted infections is high among people who engage in sexual activity for income, employment, or non-monetary items such as food, drugs and shelter[15]. A study done by Cowan et al revealed that female sex workers in Zimbabwe have a much higher risk of HIV than those in the general population and also they report high rates of symptomatic STIs which are likely to further increase the risk of contracting and transmitting HIV[10].

Various studies have shown that long distance truck drivers make up the majority of clients of sex workers especially in border towns and along major transportation routes[13, 14]. A study done on LDTD in Kenya revealed that on average they spend 4 weeks on a trip outside of Kenya, thus necessitating a prolonged absence from home and family[17]. Three quarters of these drivers reported having sex with FSW, with less than half of the participants using condoms during sex. This resulted in 50% of men reporting a history of urethritis, 26% reported a history of genital ulcer disease and 25% had serologic evidence of past or current syphilis as a result of high risk sexual behaviour[17].

Amongst LDTD and FSW, alcohol use, intravenous drug use, marijuana use and engaging in anal intercourse have been shown to be some of the common risk taking behaviours[18, 19]. Studies in India document that men who frequent community-based alcohol outlets or wine shops are more likely to engage in high risk sexual behaviour, and LDTD who consumed alcohol were 2.71 times more likely to visit a FSW than those who did not[20, 21].

Since most border towns are located in low resource areas, access to STI diagnosis is a challenge. This is against the fact that prompt recognition and treatment of STIs is paramount in controlling their transmission[38]. In countries where definitive diagnoses are difficult, the syndromic approach to management of STIs is recommended and practiced[38]. This is because HIV transmission is increased with co-existent STIs. Hence HIV public health prevention approaches must include STI preventive strategies to be effective. According to the World Health Organisation (WHO), syndromic management of STIs relies on identifying consistent groups of STI symptoms and easily recognized signs (syndromes). This will then lead to the provision of treatment that deals with the majority or most serious organisms responsible for producing a syndrome[39]. Self reporting of these syndromes by female sex workers and long distance truck drivers can give a proxy as to the prevalence of STIs in these high risk groups, in the absence of laboratory results for confirmation.

2.2. Weighting in the analysis of survey data

In order to make data collected from a survey more representative, survey weights are used. A survey weight is a value that is assigned to each case in the data file which is normally used to make statistics computed from the data to be more representative of the population the sample was drawn from[22]. It has been shown that unweighted estimators of regression coefficients may be biased if the inclusion of units in the sample is correlated with the outcome variable conditional on the explanatory variables[23]. Weighting by the reciprocals of the unit inclusion probabilities enables this bias to be corrected and regression coefficients to be estimated consistently[24].

Logistic regression is used to model binary response variables for which the response outcome for each subject is “success” or “failure”[26]. It involves fitting an equation to the data. The equation is of the form:

$$\textit{Logit}\ (p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_i X_i.$$

The goal is to find the best set of coefficients so that cases that belong to a particular category will, when using the equation, have a very high calculated probability that they will be allocated to that category. The principle by which it does so is maximum likelihood estimation which maximises the probability of getting the observed results given the fitted regression coeefficients[26]. There are two primary reasons for choosing the logistic distribution. First, from a mathematical point of view, it is an extremely flexible and easily used function, and second, it lends itself to a clinically meaningful interpretation[25]. Due to its flexibility and robustness, logistic regression can be used to model weighted survey data if the outcome variable is dichotomous. Some of the more recent improvements in logistic regression statistical software include routines to perform analyses with data obtained from complex surveys. The essential idea is to set up a function that approximates the likelihood function in the finite sampled population with a likelihood function formed from the observed sample and known sampling weights[25].

Weighted and unweighted multivariate logistic regression was used in the analysis of survey data to determine significant risk factors associated with adolescent marijuana use[28]. The results showed that the weighted model, which incorporated the complex sampling methods that were used in collecting data, was more sufficient in the analysis, leaving the authors to conclude that weights are a necessary component of the model[28]. Weighted logistic regression is an analytic technique of choice for analyzing complex survey data since it can examine the association between a categorical outcome and a number of independent variables[29]. Use of weighted logistic regression in analyzing survey data helps to account for disproportionate sampling fractions, inequalities in sampling frames, non responses and also adjusting for non probability samples like RDS and TLS[30].

Regular statistical software which is not designed to analyze survey data, analyzes data as if the data were collected using simple random sampling[40]. However, it is widely known that when surveys are conducted, a simple random sample is rarely collected. Not only is it nearly impossible to do so, but financially and statistically it is not as efficient as other sampling methods[41]. When any sampling method other than simple random sampling is used to collect data, we often need to use survey data analysis software to factor in the differences between the design that was used and simple random sampling[42]. This is due to the fact that the sampling design affects the calculation of the standard errors of the estimates. If one ignores the sampling design for example assuming simple random sampling was used when in fact another sampling design was used, the standard errors will likely be underestimated, possibly leading to results that seem to be statistically significant when in fact they are not[40].

In cases where weights were not assigned during a survey, a simple random sample can be drawn from the survey data, and probability weights can be calculated. A probability weight denotes the inverse of the probability of being included in the sample due to the sampling design[43]. The probability weight is calculated as ***N/n,*** where ***N*** is the number of elements in the population and ***n*** is the number of elements in the sample[41]. Drawing a simple random sample in STATA is a process that commences with one selecting a seed. A seed is a number that STATA begins with to generate random numbers that enables the selection of a simple random sample[44]. This number is randomly chosen from the set {0, 1, ……2 147 483 647}. Once the seed is set, the sample command is used to indicate the percentage to be sampled from the data[40]. When the sample has been drawn, probability weights are then calculated using the formula ***N/n***. If the sampling fraction (the number of elements or respondents sampled relative to the population) becomes large, a finite population correction (FPC) is used[45]. The FPC will be used in the calculation of standard error of estimates in cases where sampling without replacement is used.

Analyzing this data in STATA will require the use of the ***svy*** prefix in all the commands once the data has been set in survey mode using the ***svyset*** command[40]. This will give appropriate weighting to the data since it will consider the sampling design that has been used and in the process, getting the point

estimates and standard errors right. The default method for estimating standard errors when we specify the data to be in survey mode using the ***svyset*** command is first order Taylor linearization* which accounts for the survey design characteristics in the point estimates and variance estimation method[46]. In the non survey context, this variance estimator is called the robust variance estimator, known in STATA as the Huber/White/sandwich estimator, and it does not consider the sampling design in the estimation of point estimates and standard errors[43].

2.3. Research questions

What are the effects of including weights on the precision of regression estimates as compared to not considering weights?

Are the risk factors obtained through weighting of data and unweighting similar or do they differ?

2.4. Justification

Respondent Driven Sampling and Time Location Sampling are complex, non probabilistic survey sampling techniques. Analysis of data collected using these sampling techniques needs to consider aspects of sampling weights, clustering as well as stratification. RDSAT was created to analyze data collected through RDS, but as has been shown, the data collected failed to satisfy the assumptions for RDSAT analysis. To analyze this data, weighted logistic regression will be used since the data satisfies the assumptions for logistic regression.

Risk taking behaviours of SW and LDTD are generally unknown. Analysis of the Sero-behavioural survey data using weighted logistic regression models will yield more information with respect to risk taking behaviours of SW and LDTD which are associated with STIs.

In addition, use of weights in the analysis of this survey data that was collected using non-probabilistic sampling methods will help to show the importance or lack thereof of considering weights in the regression analysis of survey data.

* See Annexe for an explanation of Taylor linearization

2.5. Study objectives

Broad objective

To assess the effects of weighting on risk taking behaviours associated with STIs among FSW and LDTD in Beitbridge using weighted and unweighted logistic regression models.

Specific objectives

- To assess the effects of weighting on sexual and other risk behaviours associated with STIs among FSW and LDTD in Beitbridge using weighted and unweighted logistic regression models.
- To assess the effects of weighting on demographic characteristics associated with STIs in FSW and LDTD using weighted and unweighted logistic regression models
- To compare the results obtained through weighted and unweighted logistic regression models to determine the model that best explains the association between STIs and risk taking behaviours.

CHAPTER THREE

3. METHODOLOGY

3.1. Secondary Data Analysis

3.1.1. Description of application data

In this study, secondary data analysis was carried out using data from the Repeated HIV Sero-Behavioural Survey among Sex workers and Truck drivers in 3 border sites in the SADC region. This data is part of the country specific data whereby the sample size at each site had adequate power to carry out country specific analysis. Henceforth, data for FSW and LDTD in Beitbridge was used.

3.1.2. Sample size

In the original study, a total of 359 FSW and 712 LDTD were included in the study. For the purposes of using logistic regression, a minimum sample size of 50 is required for analysis. Using the Dobson[51] formula for cross sectional surveys:

$$\mathbf{n= Z_{\alpha}^2\, p\, (1-p)/d^2}$$

Where **Z** = 1.96 for **α** = 0.05, **p** is the prevalence of HIV which is 70% for FSW according to a recent study by Cowan et al[7] and **d** is a precision set at 0.05, the minimum sample size of 323 FSW was used for analysis in the weighted model.

Using the same formula for LDTD, with **p** the prevalence of HIV set at 31%[48], a minimum sample size of 328 LDTD was used for analysis in the weighted model.

In the unweighted models, a total of 359 FSW and 712 LDTD were used in the analysis.

3.1.3. Variables for secondary data analysis

For this project, the outcome variable was **STI** which was coded ‘0’ if someone never experienced signs and symptoms of an STI and ‘1’ if someone experienced signs and symptoms of an STI in the 12 months preceding the survey. The six symptoms that were considered were:

- Pain/burning sensation when urinating
- Discharge or unusual fluid coming out of the penis/vagina
- Ulcer or sore on the penis/vagina

- Ulcer or sore on the anus
- Warts on the penis/vagina
- Warts on the anus

If one responded yes to any one or more of these symptoms then they were considered to have suffered from an STI. Table 1 and 2 presents a description of some of the variables for FSW and LDTD extracted for analysis and how they were coded respectively. For a complete list of all the variables extracted for analysis, refer to annexe 3.

Table 1: Data dictionary for variables selected for secondary data analysis for FSW

	Variable description	**Variable code**	**Variable format**	**Variable name**
Outcome	Sexually transmitted infections	Ever experienced signs and symptoms of an STI in the preceding 12 months 0 = no 1= yes	Binary	STI
Study variables	Age	Age at last birthday	Numeric	AGE
	Education level	Highest level of education attained 1= at least primary 2= junior school 3= senior school	Nominal	EDULEVEL

	&above		
Alcohol intake	Ever taken drinks containing alcohol 1= no 2= yes	Binary	ALCOHOL
Drug use	Ever taken drugs other than for the purposes of medication 1= no 2= yes	Binary	DRUGS
Recent drug use	Used drugs other than for the purposes of medication in the past 12 months 1= no 2=yes	Binary	RDU
Condom use	Was a condom used the first time you had sex 1 = no 2 = yes	Binary	CONUSE
Frequency of condom use	Frequency of condom use during sex in the past 12 months	Nominal	FCONUSE

	1 = everytime 2 = almost everytime 3 = sometimes 4 = never		
Condom use for anal sex	Frequency of condom use for anal intercourse 1 = no anal sex 2 = everytime 3 = almost everytime 4 = sometime 5 = never	Nominal	ANALCON
Paying clients	Number of paying clients in the past seven days	Numeric	PAYING
Condom use with non paying partners	Was a condom used during sex with a non paying partner 1= no 2= yes	Binary	NPCON

Table 2: Data dictionary for variables selected for secondary data analysis for LDTD

	Variable description	Variable code	Variable format	Variable name
Outcome	Sexually transmitted infection	Ever experienced signs and symptoms of an STI in the preceding 12 months 0 = no 1= yes	Binary	STI
Study variables	Age	Age at last birthday	Numeric	AGE
	Length of service	Length of service as a truck driver in years	Numeric	LOSYRS
	Alcohol intake	Ever had drinks containing alcohol 1 = no 2 = yes	Binary	ALCOHOL
	Drug use	Ever taken drugs other than for the purposes of medication 1 = no 2 = yes	Binary	ETD
	Drug use for	Ever taken	Binary	DRUGSEX

sex	drugs during sex to increase pleasure 1 = no 2 = yes		
Condom use	Was a condom used during the first sexual encounter 1 = no 2 = yes	Binary	CONUSE
Sexual partners	Number of regular female sexual partners	Numeric	RFP
Sexual partners	Number of non-regular female sexual partners	Numeric	NRFP
Frequency of condom use	Frequency of condom use with non-regular female sexual partners 1 = everytime 2 = almost everytime 3 = sometimes 4 = never	Nominal	NRPCONUSE

5 = don’t know
6 = no non
regular
partner
7 = no vaginal
intercourse

3.2. Methodology of analysis

3.2.1. Data extraction

The data was already in STATA format when it was accessed by the researcher. In the original dataset, variables were grouped into eight sections which were:

- Background characteristics
- Alcohol and drug use
- Sexual history
- Male condom
- STIs
- Knowledge and attitudes about HIV/AIDS
- Access to services
- Income and expenditure.

For secondary data analysis, variables were extracted from 4 groups which were deemed to determine risk taking behaviours namely: background characteristics, alcohol and drug use, sexual history and male condom. The outcome variable was extracted from the STI group. After extraction, the datasets for FSW and LDTD were kept separately.

3.2.2. Data management

The data had been cleaned and coded by the original researcher. However, for the sake of this project data was re-coded as shown in table 1 and 2 specifically for this analysis. The original data quality was high so much so that there were very few cases with missing values. These were treated as missing completely at random since their being missing was unrelated to actual values of the missing

data. There were no missing values for the variables that were used to determine the outcome variable.

3.2.3. Statistical methods

Weighted and unweighted logistic regression models were used to determine risk taking behaviours associated with STIs in both female sex workers and long distance truck drivers. In order to use logistic regression, the outcome variable for both FSW and LDTD which was “Ever suffered from STI symptoms in the past 12 months” was coded ‘0’ if the response was “no” and ‘1’ if the response was “yes”. Logistic regression is a type of regression which is used when the dependent or outcome variable is binary, discrete or categorical and the predictor or independent variables are of any kind. It is particularly useful in health sciences as the binary outcome is often the presence or absence of some health condition or disease.

Logistic regression uses the logit transformation to predict group membership based on several covariates irrespective of their underlying distribution thus it avoids predicting negative probabilities of group membership[26]. It is especially important when the outcome has a non-linear (sigmoidal) relationship with the independent variables. Logistic regression analysis is based on the calculation of the odds of an outcome, which is the ratio of the probability of having an outcome or belonging to one group divided by the probability of not having the outcome or not belonging to that group.

Assumptions of logistic regression that have to be met include[25, 26]:

- Linearity in the logit for continuous variables
- Independence of error terms
- Absence of multicolinearity
- Lack of strongly influential outliers.

3.2.4. Statistical analysis

In the weighted and unweighted analysis FSW and LDTD’s demographic characteristics were described using means and standard deviations for

continuous, normal variables. Medians and inter-quartile ranges were used for non-normal continuous variables. Post estimation commands were used to get standard deviations for continuous variables in the weighted analysis since by default the output reports mean and 95% confidence interval. Continuous variables were tested for normality using histograms. Categorical variables were described using frequencies and percentages. Odds ratios were used to describe the association between the predictor variables and the dependent variable.

i. Unweighted models

Forward selection was used to build the unweighted logistic regression model for FSW and LDTD. For both groups, the outcome variable was STI. A pool of potential predictor variables were selected from the dataset which included sexual factors, alcohol and drug use factors as well as behavioural factors. Univariate analyses were run on all the predictor variables that had been selected. A cut off p – value of 0.25 was used to determine predictor variables that were to be considered for the multivariate model as is recommended by Hosmer and Lemeshow[22]. They argue that use of traditional cut off p – value of 0.05 as a screening tool for considering candidate variables for the multivariable model often fails to identify variables that are known to be important. A problem that is usually common with the univariate approach to model selection is that it ignores the possibility that a collection of variables, each of which is weakly associated with the outcome can become an important predictor of the outcome when taken together[22]. In cases where this is thought to be a possibility, then a significance level large enough to allow the suspected variables to be candidates for inclusion into the multivariable model is used, hence 0.25. For categorical variables with *k-1* levels, if any one of the levels had a p – value greater than 0.25 then the variable was not considered for the multivariate model.

Multivariate model building commenced with a no relationship model (i.e. the intercept only model) and its log likelihood (L_0) was evaluated. A predictor variable with the lowest p – value was then added to this intercept only model. Separately, each of the predictor variables selected for the multivariate model were added starting with ones with lower p – values. Predictors which became insignificant were eliminated from the model and only those variables which were

significant were retained. The model with two significant variables with the lowest p – value became the interim model. A third variable was added to this interim model starting with the one with the lowest p – value and the same process was repeated until all the predictors had been included in the multivariate model. Only those predictors that were significant were retained in the model.

Unweighted model diagnostics

In order to check that the logit function was the correct function to use and that all relevant predictor variables were included in the model, the **linktest** command was used. This was to detect a specification error. The idea behind a **linktest** is that if the model is properly specified, there should not be any additional predictors that are statistically significant except by chance. The **linktest** uses the linear predicted value (_**hat**) and linear predicted squared (_**hatsq**) to rebuild the model. Ideally, _**hat** should be statistically significant while _**hatsq** should be insignificant if the model is correctly specified.

Assessing the goodness of fit of the model was done through the use of the log likelihood chi-square, pseudo R-square and the Hosmer and Lemeshow goodness of fit. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were also used to asses model fit. Collinearity was assessed using the Variance Inflation Factor (VIF) and the tolerance. Pearson residuals and deviance residuals were used to measure the deviations between observed and fitted values. The Pregibon leverage was used to measure the leverage of an observation. Another statistic called Pregibon's dbeta, which is similar to Cook's distance in Ordinary Linear regression was used to obtain summary information of influence on parameter estimates of each individual observation. Significance of parameters was tested using individual Wald tests[22, 23]. Likelihood ratio tests were used to test model significance.

Classification tables were used to determine the sensitivity and specificity of the model. Higher specificity and sensitivity indicating a better fit of the model. A scatter plot of the sensitivity and one minus specificity provided an ROC (Receiver Operating Characteristic) curve. The overall fit of the model is determined by the area under the ROC curve.

*ii. **Weighted models***

Since weights were not assigned during data collection, it was necessary to generate weights that were to be used in the weighted models. This required drawing random samples from the data that was already available. The calculated sample size guided the drawing of random samples for FSW and LDTD. In STATA this was done by first running the command **count**. This would show the original sample size in the data set. The second step was to set the seed, which was achieved by running the command **set seed 1003002849** as is recommended by the UCLA Statistical Consultancy Group[40]. After setting the seed, the proportion of the sample to be drawn was then specified (90% for FSW and 46% for LDTD) based on the sample size calculated for secondary data analysis. The **count** command was run again after specifying the sample proportion to show the new sample size (323 FSW, 328 LDTD). Probability weights (pweight) were then generated for FSW and LDTD and these were calculated as ***N/n***. A finite population correction (FPC) was generated which is usually equated to the original sample size and STATA will make the necessary calculations to obtain the correct FPC.

To set the data into survey mode, the command ***svyset [pweight=pw], fpc(fpc)*** was run. This command assigns probability weights (pweights) and calculates the finite population correction (fpc). After this command, the data is now in survey mode and all commands to be executed are to be prefixed by the survey prefix ***svy***. To display the information regarding the sampling plan the command ***svydes*** was run.

The same model building technique that was used in the unweighted models was used for the weighted models. The only difference being that for commands in the weighted models, the prefix ***svy*** was used before the ***logistic*** command. This was meant to factor in the new sampling design as well as the weights. Univariate and multivariate models were run as in the unweighted models. The same criterion for inclusion into the multivariate analysis for predictor variables that was used in the unweighted models was applied.

Weighted models diagnostics

Unlike in the unweighted models, goodness of fit test for weighted models uses the survey logistic goodness of fit (***svylogitgof)*** command to compute model diagnostics. This is because after using the survey (***svy)*** command, traditional Hosmer-Lemeshow goodness of fit tests are not available. For survey data Archer and Lemeshow recommend the use of the *F*- adjusted mean residual test[31]. This was used to assess model goodness of fit. Collinearity was checked using the VIF and the tolerance.

3.3. Ethical considerations

Approval to carry out the secondary data analysis was sought and granted by the Joint Research Ethics Committee (approval letter in Appendix A1). Permission to use the data was granted by the principal investigator, Prof S Rusakaniko (letter of authorization in Appendix A2)

CHAPTER FOUR

4. Results

4.1. Demographic characteristics

Female Sex workers

In the unweighted analysis, a total of 359 FSW were included in the study. The mean (SD) age was 27.9 (6.8) years. Most of the FSW (47.1%) had attained at least a senior school education and 64.1% were non-Catholic Christians.

In the weighted analysis, a total of 323 FSW were included in the study. The mean (SD) age was 27.8 (6.7) years. A comparison of respondent characteristics for the weighted and unweighted analysis is shown in table 3. From the table, it is evident that demographic variables are not significantly different whether weights are included or not.

Table 3: Demographic characteristics of FSW

Variable	Unweighted (n=359) Proportion (%)	Weighted (n=323) Proportion (%)	p-value
Age$_{\text{(mean, SD)}}$	27.9 (6.8)	27.8 (6.7)	0.393
Education level			
At least primary	18.9	20.2	0.334
Junior school	34.0	30.3	0.149
Senior school	47.1	49.5	0.263
&over			
Religion	19.8	19.2	0.417
No religion	0.6	0.6	0.501
Traditional	14.2	14.2	0.504
Roman Catholics	64.1	64.7	0.439
Non-Catholics			
Islam	1.1	0.9	0.389

Employment status			
Not employed	84.4	84.2	0.466
Part time employed	12.81	13.0	0.469
Full time employed	2.5	2.8	0.403
Marital status			
Single(never married)	34.8	34.7	0.487
Living together	0.6	0.6	0.500
Married	1.7	1.6	0.493
Divorced/Separated	49.6	49.5	0.493
Widowed	13.4	13.6	0.465

Long distance truck drivers

A total of 712 LDTD were included in the unweighted analysis. The mean age (SD) was 38.4 (8.2) years. Most of the truck drivers had attained a senior school education (79.8%).The majority of the drivers were married (90.5%).

In the weighted analysis, a total of 328 LDTD were included in the analysis after re-calculating the sample size. The mean age (SD) was 39.3 (8.6) years. The majority of the truck drivers were non-catholic Christians (76.7%) and were married (90.8%).

Weighting does not have an effect on demographics as shown in table 4. There is no statistically significant difference in estimates for demographics in the weighted and unweighted analysis.

Table 4: Demographic characteristics of LDTD

Variable	Unweighted (n=712) Proportion (%)	Weighted (n=328) Proportion (%)	p-value
Age$_{(mean, SD)}$	38.4 (8.2)	38.8 (8.4)	0.338
Education level			
At least primary	4.6	5.2	0.343
Junior school	9.8	9.5	0.438
Senior school	79.8	79.0	0.388
Tertiary	5.8	6.4	0.362
Religion			
No religion	10.8	10.1	0.369
Traditional	1.0	0.6	0.275
Roman catholic	12.4	11.0	0.264
Non-Catholics	73.3	76.2	0.172
Islam	2.3	1.8	0.303
Marital status			
Single	6.2	5.2	0.279
Living together	1.1	0.9	0.376
Married	90.5	90.6	0.484
Divorced	1.5	1.8	0.373
Widowed	0.7	1.5	0.119

4.2. Demographic variables associated with STIs

Female Sex workers

An evaluation of demographic variables showed that age was independently significantly associated with having suffered an STI within the past 12 months in the weighted analysis [OR; 95%CI] (1.04; 1.03-1.05). For every year increase in age for female sex workers, the odds of suffering from an STI were increased by 3.8% in the weighted analysis. However, in the unweighted analysis, age was not significantly associated with suffering from an STI. However, despite other demographic variables being insignificant, inclusion of

weights improves the precision of estimates as shown by the narrow confidence intervals in the weighted analysis as compared to those in the unweighted analysis for the same variables. Table 5 summarises demographic variables associated with STIs for female sex workers.

Table 5 : Multivariable logistic regression for demographic variables independently associated with STIs for FSW

Variable	Weighted OR(95% CI)	*S.E**	Unweighted OR (95% CI)	*S.E**
Age	1.04 (1.03 – 1.05)	0.01	1.027 (1.00 - 1.06)	0.02
Junior school education	1.17 (0.95 – 1.43)	0.12	1.10 (0.60 – 2.00)	0.34
Senior school education &over	0.70 (0.58 – 0.83)	0.06	0.66 (0.38 – 1.16)	0.19
Shona	2.16 (1.71 – 2.72)	0.25	2.29 (1.13 – 4.68)	0.83
Venda	1.08 (0.78 – 1.50)	0.18	1.22 (0.45 – 3.25)	0.61
Other ethnicity	3.15 (2.24 – 4.41)	0.54	0.36 (1.32 – 10.11)	1.90
Part time employed	1.59 (1.28 – 1.97)	0.17	1.53 (0.81 – 2.87)	0.49
Full time employed	1.22 (0.79 – 1.88)	0.27	1.23 (0.32 – 4.65)	0.83

**Standard error*

Long distance truck drivers

None of the demographic variables for LDTD were independently significantly associated with having suffered an STI within the preceding 12 months both in the weighted and unweighted analysis. Table 6 shows demographic variables associated with STIs for LDTD. From the table, it is shown that inclusion of weights results in narrow confidence intervals and hence higher precision of estimates.

Table 6 : Multivariable logistic regression for demographic variables independently associated with STIs for LDTD

Variable	Weighted OR (95% CI)	*S.E**	Unweighted OR (95% CI)	*S.E**
Age	0.99 (0.96 – 1.02)	0.01	0.99 (0.97 – 1.02)	0.01
Senior school education	0.49 (0.21 – 1.17)	0.22	1.17 (0.40 – 3.42)	0.64
Tertiary education	0.34 (0.08 – 1.33)	0.24	1.49 (0.40 – 5.61)	1.01

**Standard error*

4.3. Behavioural factors associated with STIs

Female Sex workers

Univariate analysis

Table 7 shows results of the weighted and unweighted univariate analysis performed to investigate the independent association between the predictor variables and the dependent variable. The odds ratios are higher in the weighted analysis as compared to the unweighted analysis. Furthermore, the confidence intervals in the unweighted analysis are wider as compared to those in the weighted analysis for the same variables. In addition, more variables were available for selection into the multivariate analysis in the weighted analysis using a cut off p-value of 0.25.

Table 7 : Univariate weighted and unweighted logistic regression analysis for factors associated with STIs for FSW

Variable	Weighted OR(95%CI)	p-value	Unweighted OR(95%CI)	p-value
Alcohol use	2.55 (2.20 – 2.96)	<0.001	2.37 (1.53 – 3.66)	<0.001
Ever taken drugs	2.18 (1.81 – 2.63)	0.089	1.75 (1.04 – 2.95)	0.038

Recent drug use	2.15 (1.77 – 2.61)	<0.001	1.87 (1.08 – 3.23)	0.027
Used other drugs	0.37 (0.11 – 1.21)	0.104	0.37 (0.12 – 1.11)	0.083
Used drugs to increase sexual pleasure	2.10 (1.77 – 2.49)	<0.001	1.95 (1.18 – 3.23)	0.011
Used Viagra to increase sexual pleasure	4.33 (1.56 – 2.08)	0.011	4.8 (1.60 – 14.44)	0.011
Used other drugs to increase sexual pleasure	0.27 (0.10 – 0.72)	0.013	0.30 (0.11 – 0.81)	0.023
Age at first sex	0.94 (0.91 – 0.96)	<0.001	0.93 (0.86 – 1.01)	0.088
Used a condom during the first sexual encounter	0.46 (0.38 – 0.54)	<0.001	0.48 (0.28 – 0.82)	0.009
Number of paying clients last 7 days	1.01 (1.00 – 1.02)	0.010	1.02 (1.00 – 1.04)	0.101
Number of clients on the last day worked	1.09 (1.05 – 1.13)	<0.001	1.08 (0.98 – 1.19)	0.115
Had oral sex with paying client	1.38 (1.07 – 1.78)	0.012	1.58 (0.72 – 3.44)	0.252*
Paid extra not to use a condom	2.19 (1.90 – 2.52)	<0.001	2.38 (1.56 – 3.63)	<0.001
Used a condom with non paying client	0.72 (0.52 – 0.99)	0.041	0.67 (0.41 – 1.12)	0.133
Age at first drink	1.03 (1.00 – 1.07)	0.090	1.02 (0.97 – 1.08)	0.382*
Cannabis use	2.08 (0.70 – 6.22)	0.181	1.75 (0.63 – 4.89)	0.284*
Age first received money for sex	1.02 (1.00 – 1.03)	0.011	1.00 (0.97 – 1.04)	0.889*

Use condom everytime for anal sex	0.86 (0.68 – 1.09)	0.213	1.09 (0.55 – 2.13)	0.811*
Use condom almost everytime for anal sex	1.63 (1.02 – 2.58)	0.036	1.55 (0.43 – 5.60)	0.514*
Sometimes use condom for anal sex	1.55 (1.21 –1.97)	<0.001	1.82 (0.89 – 3.75)	0.100
Never use condom for anal sex	2.45 (1.44 – 4.14)	<0.001	3.09 (0.61–15.60)	0.169
Had handshake sex with paying client	0.87 (.073 – 1.04)	0.133	0.92 (0.53 – 1.60)	0.755*
Had intercrural sex with paying client	0.60 (0.44 – 0.80)	<0.001	0.63 (0.26 – 1.50)	0.300*
Had anal sex with paying client	1.30 (0.90 – 1.90)	0.161	1.09 (0.36 – 3.31)	0.880*
Client suggested condom use	1.64 (1.11 – 2.42)	0.010	1.45 (0.59 – 3.56)	0.411*
Joint decision to use condom	2.19 (1.54 – 3.11)	<0.001	2.18 (0.95 – 5.02)	0.073
Number of regular non paying clients	0.98 (0.97 – 0.99)	0.011	0.98 (0.95 – 1.02)	0.567*

**Not significant for inclusion into the multivariate analysis*

Multivariate analysis

In the unweighted multivariate analysis alcohol use, used drugs to enhance sexual pleasure, used condom during the first sexual encounter and being paid extra not to use a condom were significantly associated with STI.

From the model, the odds of suffering from an STI by a female sex worker who used drugs to increase sexual pleasure were increased by 78% adjusting for other factors. Likewise, the odds of suffering from an STI by a female sex worker who took alcohol were increased by 140% adjusting for other factors.

Female sex workers who used a condom on their first sexual encounter had a 61% reduction in the odds of suffering from an STI adjusting for other factors.

The weighted model has almost the same parameters as the unweighted model. However, the odds ratios are much higher in the weighted model than in the unweighted model. The odds of suffering from an STI by a female sex worker who took alcohol were increased by 146% adjusting for other factors in the weighted model. In the unweighted model, the same odds were increased by 140% adjusting for other factors.

Table 8 compares the odds ratios, the 95% confidence interval and the standard errors of the parameters in the weighted and unweighted models. From the table, it is shown that the odds ratios are higher, the confidence intervals are narrow and the standard errors are smaller in the weighted model than in the unweighted model.

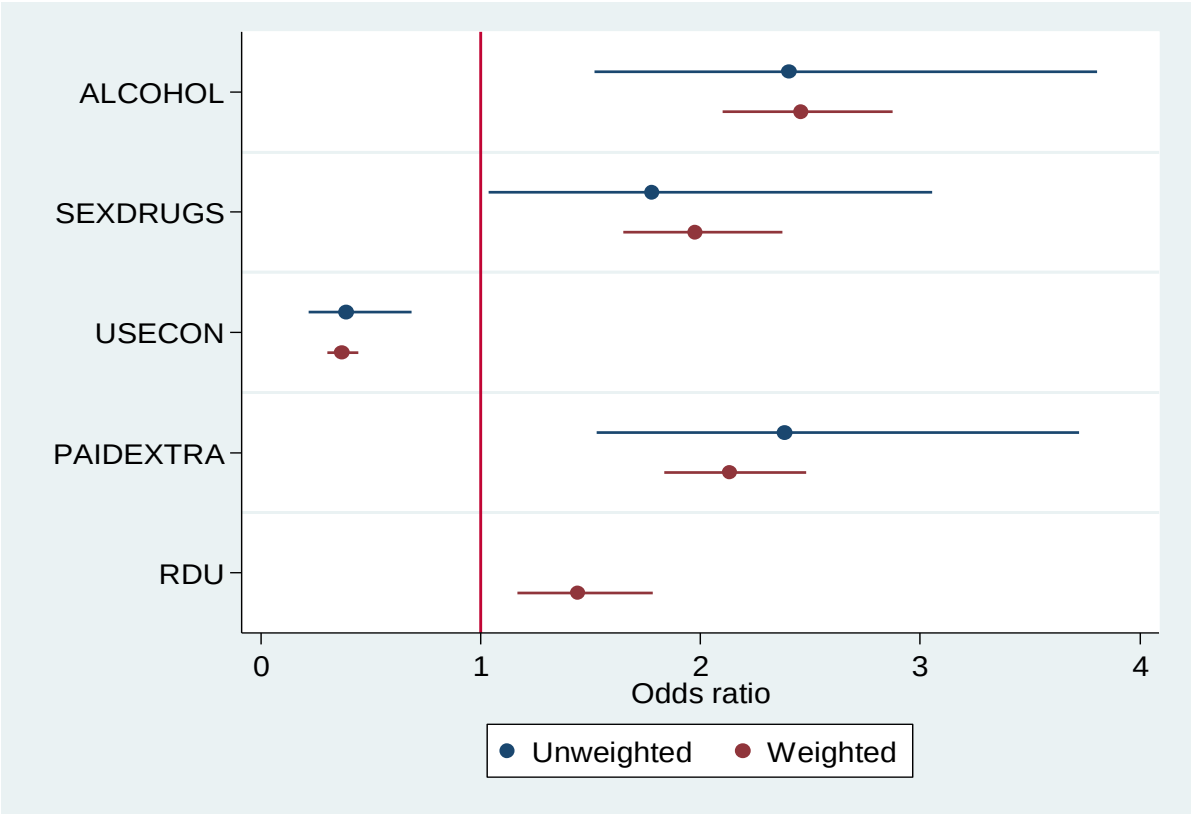
Table 8 : Multivariable weighted and unweighted logistic regression analysis for factors independently associated with STIs (FSW)

Variable	Weighted OR(95% CI)	*S.E*	Unweighted OR(95% CI)	*S.E***
Alcohol use	2.46 (2.10 – 2.87)	0.20	2.40 (1.52 – 3.80)	0.56
Used drugs to increase sexual pleasure	1.98 (1.65 – 2.37)	0.18	1.78 (1.04 – 3.06)	0.49
Used a condom during the first sexual encounter	0.37 (0.30 – 0.44)	0.04	0.39 (0.22 – 0.69)	0.11
Paid extra not to use a condom	2.13 (1.84 – 2.48)	0.16	2.38 (1.53 – 3.72)	0.54
Recent drug use	1.44 (1.17 – 1.78)	0.16	1.35 (0.74 – 2.45)	0.41*

**Not significant, **Standard error*

A comparison of the coefficients is shown graphically in figure 1 for the two models. The plot shows the coefficient and its associated 95% confidence interval. It can be seen that for the weighted model, the 95% confidence intervals are narrow as compared to the unweighted model thus implying greater precision of the weighted model.

Figure 1: Coefficient plots for the weighted and unweighted models (FSW)



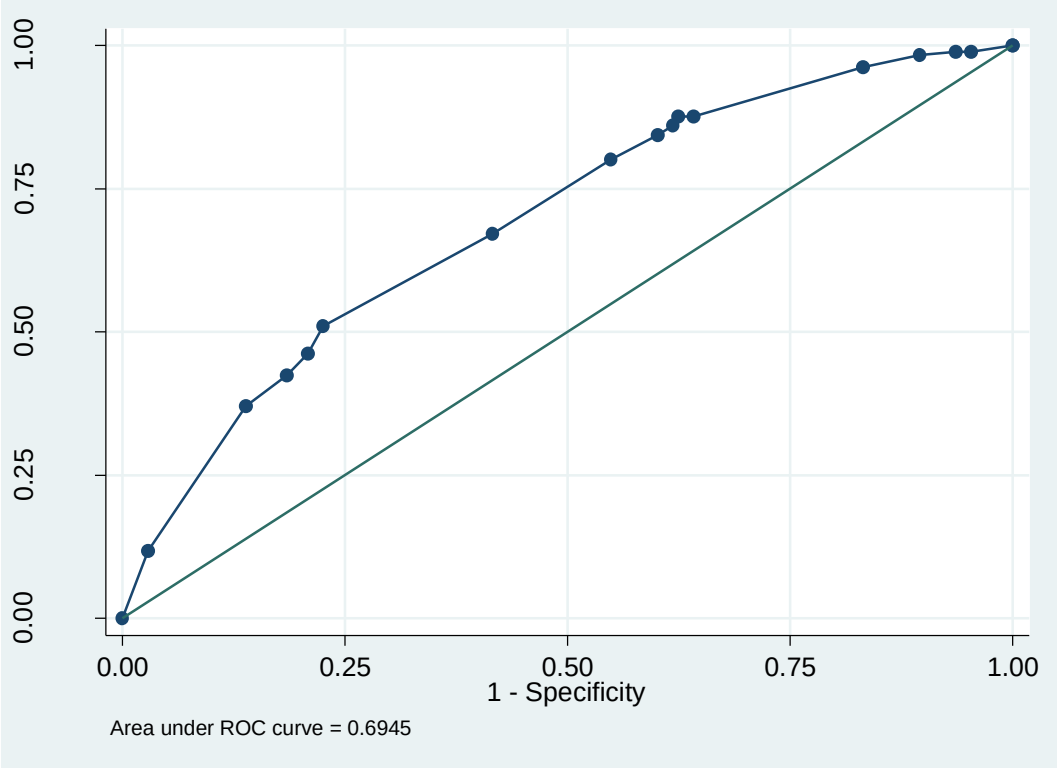
Goodness of fit tests

There is no evidence of lack of fit for the weighted model (Archer & Lemeshow F-adjusted test statistic (9, 315) = 1.92; p = 0.06). The linktest suggests that the model is correctly specified (_hatsq = 0.18) while individual Wald tests indicate that the parameters in the model are significantly not equal to zero. In the absence of the likelihood ratio test, using the F-test to determine model significance shows that the model is statistically significant [F (5, 318) = 73.44; p <0.001].

The likelihood ratio test of the unweighted model was significant (p<0.001) indicating overall significance of the model. Individual Wald tests were also significant indicating that the parameters in the model are significantly not equal to zero. The Hosmer and Lemeshow goodness of fit test was insignificant (p = 0.99) indicating that model fit is good. The linktest was insignificant (_hatsq = 0.80) which means that the model is correctly specified. McFadden's pseudo R^2 was 0.09 which is close to zero thus indicating a poor predictive value of the model.

The unweighted model had a sensitivity of 67.2% and a specificity of 58.4% using a cut off of 0.5 which represents good model classification. The area under the ROC curve points to a good predictive power of the model.

Figure 2: Receiver Operating Curve (ROC) for the unweighted model (FSW)



(Key: The 45 degree line is called the chance line. It represents a discriminating power of the model which is not better than chance if all the plots lie on the line. The area under the curve represents the accuracy of the model, thus 69.5% accuracy for this model)

Long distance truck drivers

Univariate analysis

Table 9 below shows the results of the weighted and unweighted univariate analysis to determine independent associations between predictor variables and the dependent variable. Most of the odds ratios from the weighted analysis are generally higher compared to those obtained in the unweighted analysis and the confidence intervals are mostly narrow in the weighted analysis.

Table 9: Univariate weighted and unweighted logistic regression analysis for factors associated with STIs (LDTD)

Variable	Weighted OR(95% CI)	p-value	Unweighted OR (95% CI)	p-value
Away from home for more than a month	2.92 (1.61 – 5.30)	<0.001	2.70 (1.59 – 4.57)	<0.001
Length away from home	0.93 (0.83 – 1.04)	0.208	0.92 (0.82 – 1.04)	0.172
Alcohol use	1.50 (0.92 – 2.44)	0.102	1.67 (1.08 – 2.58)	0.022
Age at first drink	0.92 (0.86 – 0.99)	0.042	0.94 (0.89 – 0.99)	0.016
Ever taken any drugs	4.29 (2.44 – 7.55)	<0.001	3.27 (1.93 – 5.55)	<0.001
Used drugs in the past 12 months	3.27 (1.62 – 6.50)	0.001	3.52 (1.89 – 6.58)	<0.001
Used cannabis in the past 12 months	1 (omitted)	_	5.63 (1.11 – 28.43)	0.037
Used cocaine in the past 12 months	1(omitted)	_	2.64 (0.70 – 9.94)	0.152
Used drugs to enhance sexual	2.05 (1.20 – 3.50)	0.008	1.42 (0.88 – 2.29)	0.154

pleasure				
Intravenous drug use	1(omitted)	_	3.22 (0.58 – 17.80)	0.181
Age at first sex	0.92 (0.84 – 0.99)	0.041	0.94 (0.89 – 1.01)	0.074
Regular female partner	1.50 (1.22 – 1.86)	<0.001	1.25 (1.04 – 1.50)	0.017
Non regular female partner	1.29 (1.16 – 1.42)	<0.001	1.17 (1.07 – 1.27)	0.001
Used a male condom in the past 12 months	3.86 (2.06 – 7.20)	<0.001	3.78 (2.19 – 6.34)	<0.001
Always carry condoms	2.50 (1.51 – 4.15)	<0.001	2.30 (1.47 – 3.60)	<0.001
Commercial female client	1.15 (1.09 – 1.21)	0.001	1.00 (0.99 – 1.01)	0.322*

**Not significant*

Multivariate analysis

In the unweighted model, the odds of suffering from an STI within the past 12 months were increased by 149% for a truck driver who was away from home for more than one month adjusting for other factors. Similarly, the odds of suffering from an STI within the past 12 months were increased by 153.4% for a truck driver who ever used drugs adjusted for other factors. Likewise, the odds of suffering from an STI within the past 12 months were increased by 10.3% for every non-regular female sexual partner a truck driver had adjusted for other factors. Using a male condom within the last 12 months increased the odds of suffering from an STI by 215.4% adjusting for other factors.

The weighted and unweighted models had two common parameters (ETD and AFHM). The odds of suffering from an STI as a result of ever taking drugs were higher in the weighted model than in the unweighted model adjusting for other

factors. Interestingly, the odds of suffering from an STI as a result of being away from home were higher in the unweighted model than in the weighted model, adjusting for other factors. However, the weighted model is more biologically plausible compared to the unweighted model. Table 10 compares the variables in the multivariate analysis for the weighted and unweighted analysis.

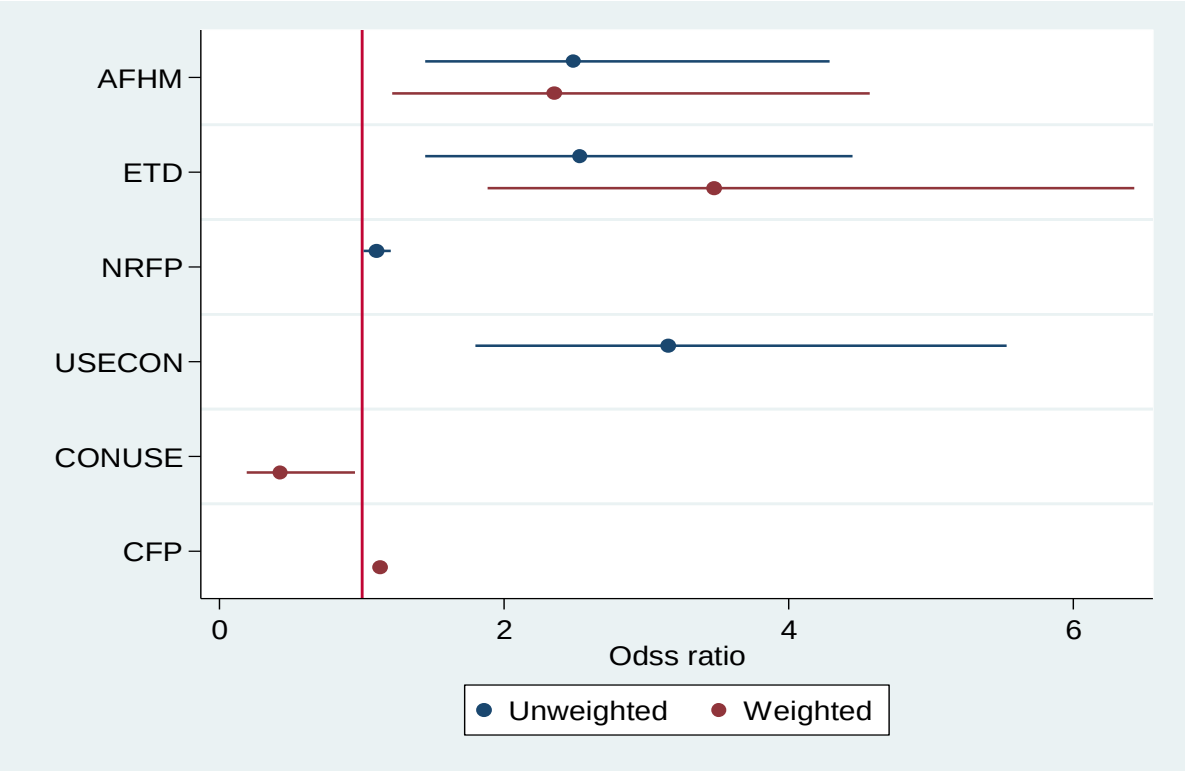
Table 10: Multivariable weighted and unweighted logistic regression analysis for factors independently associated with STIs (LDTD)

Variable	Weighted OR(95% CI)	*S.E*	Unweighted	*S.E***
Ever taken any drugs	3.48 (1.88 – 6.42)	1.09	2.53 (1.44 – 4.45)	0.69
Away from home for more than a month	2.35 (1.21 – 4.57)	3.15	2.49 (1.45 – 4.29)	0.73
Commercial female client	1.13 (1.08 – 1.18)	0.03	0.99* (0.98 – 1.00)	0.01
Non regular female partner	1.05* (0.88 – 1.25)	0.09	1.10 (1.01 – 1.21)	0.05
Used a male condom in the past 12 months	2.55 (1.31 – 4.96)	0.86	3.15 (1.80 – 5.53)	0.90
Used a condom on the first sexual encounter	0.43 (0.19 – 0.95)	0.17	1.06* (0.61 – 1.82)	0.29

**Not significant, **Standard error*

Comparing the coefficient plots of the two models (figure 3); the weighted model shows desirable characteristics since it has a better precision as evidenced by the relatively narrow confidence intervals as compared to the unweighted analysis.

Figure 3: Coefficient plot for the weighted and unweighted models (LDTD)



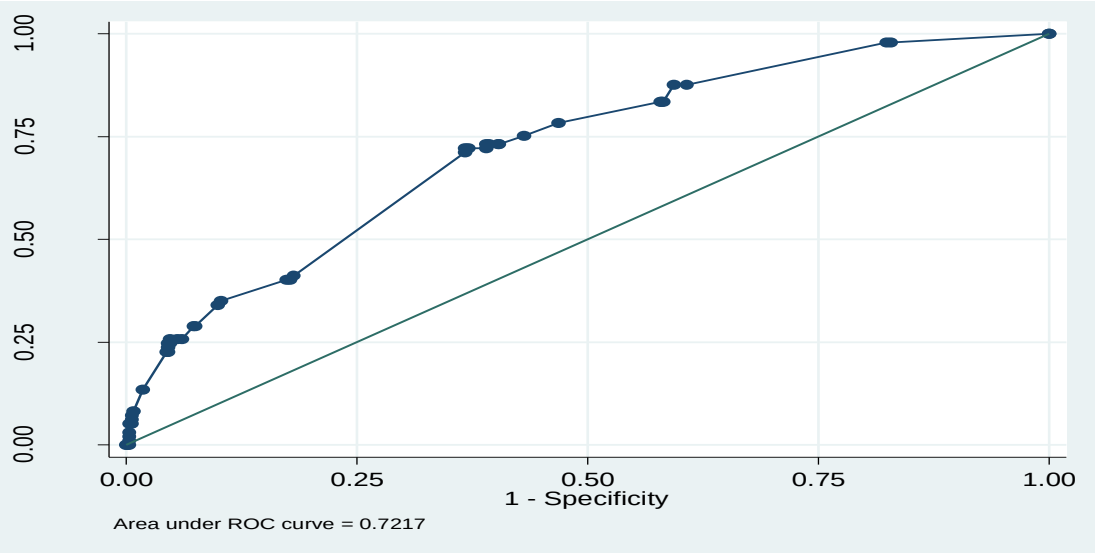
Goodness of fit tests

The model fit of the weighted model was good (Archer & Lemeshow F-adjusted test statistic (5, 317) = 0.37; p = 0.87). The linktest suggests the model is correctly specified (_hatsq = 0.30) while adjusted Wald tests suggests that the parameters in the model are significantly not equal to zero. The global F-test suggests that the model is statistically significant [F (2, 320) = 28.36; p<0.001].

The likelihood ratio test of the unweighted model was significant (p<0.001) indicating overall significance of the model. Individual Wald tests were significant indicating that the parameters in the model were significantly not equal to zero. The Hosmer and Lemeshow goodness of fit chi square statistic was insignificant (p = 0.68) indicating that the model fit is good. The linktest was insignificant (_hatsq = 0.34) which implies that the specification is good. McFadden's pseudo R^2 was 0.104 which is close to zero, thus the model has a poor predictive value.

For a cut off point of 0.5, the model has a poor sensitivity (3.09%) and a very good specificity (99.7%). The area under the ROC curve (72.17%) indicates a good predictive power of the model.

Figure 4: Receiver Operating Curve (ROC) for the unweighted model (LDTD)



(Key: The 45 degree line is called the chance line. It represents a discriminating power of the model which is not better than chance if all the plots lie on the line. The area under the curve represents the accuracy of the model, thus 72.17% accuracy for this model)

CHAPTER FIVE

5. DISCUSSION AND CONCLUSION

5.1. Discussion

5.1.1. Introduction

Weighting in the regression analysis of survey data collected through non-probabilistic sampling methods is meant to adjust for non-equal probability of selection, disproportionate sampling and non-response. In essence, sampling weights weigh the sample data to correct for disproportionality of the sample with respect to the target population of interest[34]. Theoretically, weighting improves the precision of estimates in a regression model[33]. According to Kish, weights play a pivotal role in two different aspects of the modelling process which are:

- They can be used to test and protect against nonignorable sampling designs which could cause selection bias.
- They can be used to protect against misspecification of the model holding in the population[30].

Studies have shown that for weighting to improve the precision of estimates, there should be a relationship between the outcome being measured and the probability of selection into the survey [35, 36].

5.1.2. Key findings

This study has shown that weighting in the regression analysis of survey data improves the precision of the estimates. In the multivariate unweighted model for female sex workers, only four predictor variables were included. In the multivariate weighted model, the number of predictor variables rose to five including the four which initially made up the unweighted model. The precision of the estimates in the weighted model were superior to that of the unweighted model. The confidence intervals in the weighted models were narrow and the standard errors were smaller.

The findings from this study show that alcohol use, use of drugs to increase sexual pleasure, not using a condom on the first sexual encounter and being paid extra money so as not to use a condom are significant risk taking behaviours of

FSW associated with suffering from STI. These variables were common in the weighted as well as the unweighted model. However, they were more significant in the weighted model than the unweighted model. In addition to the four risk taking behaviours, recent drug use was found to be another risk taking behaviour associated with suffering from an STI in the weighted model. This shows the importance of considering the sampling design when analysing data collected through non-probabilistic sampling methods since variables that are not significant in the unweighted analysis may become significant in the weighted analysis.

For LDTD, inclusion of weights resulted in a model which was almost similar to the one obtained in the unweighted model. Inclusion of weights failed to produce a pronounced effect on the parameters in the model as was the case with FSW. The coefficients in the two models were equally comparable. The confidence intervals and standard errors were almost similar. Parameters common in both models were being away from home for more than one month and ever using drugs. Comparing the two models, the weighted model is biologically plausible since in the unweighted model, the use of condoms is a risk factor for suffering from an STI when it is supposed to be protective. The reason behind a variable which is usually protective becoming a risk factor is explained by Pfefferman. In their study they found out that when the logistic model is estimated ignoring the study design (thus assuming simple random sampling), the estimate of the intercept term will have a large bias which will have a deteriorating effect on the estimated coefficients[37].

When it comes to demographics, inclusion of weights was shown not to be necessary. There was no statistically significant difference in the estimates obtained after including weights and those obtained in the unweighted analysis. This is because simple random sampling was used to draw the samples used for the weighted analysis. Thus this sample is representative of the population it was drawn from. This population served as the sample in the unweighted analysis.

From the four models generated, the weighted models were far much better in terms of precision and biological plausibility as compared to the unweighted models. As a direct result of including weights, the standard errors became

smaller, the confidence intervals became narrow and the odds ratios were higher as compared to the unweighted models especially for FSW. In as much as there was no much change in terms of standard errors and confidence intervals for LDTD weighted model, weighting resulted in a model with biological plausibility. Thus weighted models were the best models to describe the association between STIs and risk taking behaviours for FSW and LDTD in Beitbridge.

5.2. Conclusion

The results of this study have shown the importance of including weights in the regression analysis of survey data collected through non-probabilistic sampling methods. Effects of weighting in the analysis of survey data that have been shown as a result of this study are:

- Inclusion of weights improves the precision of the estimates since the standard errors become smaller and the confidence interval becomes narrow.
- Weighting reduces selection bias since analysis will be performed on a representative sample.
- Inclusion of weights reduces variances through the use of Taylor linearization for variance estimation.
- Weights increase the predictive power of models.

This study has shown that weights improve the precision of estimates in regression analysis of survey data collected through non-probabilistic sampling methods. It has also shown that weights are not necessary in the analysis of demographic variables. Furthermore, weighted models were preferred to unweighted models since they have a higher predictive power.

Therefore, in conclusion, it has been shown that weights are important and should always be considered when analysing survey data collected through the use of non-probabilistic sampling methods.

5.3. Limitations

Results of this analysis can only be inferred to FSW and LDTD in Beitbridge at the time of the survey.

In as much as weighted models have been shown to be desirable, there are few diagnostics to validate the goodness of fit of the model. Ordinary logistic regression uses the Hosmer and Lemeshow goodness of fit test[25] to assess model goodness of fit. This test is quite robust in its evaluation of the model. Other tests include determining classification of the model, identifying influential observations, identifying outliers as well as determining sensitivity and specificity of the model. Thus the assumptions of logistic regression are easily assessed in the unweighted analysis.

In the weighted model, instead of the Hosmer and Lemeshow goodness of fit test, there is the Archer and Lemeshow test[34]. This test is not as robust as the Hosmer& Lemeshow test. Significant variables in the multivariate model were dropped so as to achieve fitness of the model. Moreover, diagnostic tests like identifying influential observations, determining sensitivity and specificity of the model as well as determining the predictive power of the model are not available in the weighted models. In a study by Baker, they found out that even though weighted models satisfy the assumptions of logistic regression, there are few model diagnostics that can be done on weighted models due to a lack of appropriate statistical diagnostics in STATA[28]. Therefore, results from weighted models should always be interpreted with caution.

6. REFERENCES

1. McCreesh N, Frost S, Seeley J, et al. Evaluation of respondent driven sampling. Epidemiology. Jan 2012; 23(1): 138-147
2. Muhib FB, Lin LS, Stueve A et al. A venue-based method for sampling hard-to-reach populations. Public Health Report 2001; 116(1): 216-222
3. Magnani R, Sabin K, Saidel T and Heckathorn D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. AIDS 2005; 19(2): S67-S72.
4. McCree H D, Cosgrove S, Stratford D et al. Sexual and Drug use risk behaviours of long-haul truck drivers and their commercial sex contacts in New Mexico. Public Health Reports. 2010;125:52-60
5. Morris CN and Ferguson AG. Estimation of the sexual transmission of HIV in Kenya and Uganda on the trans-Africa highway: the continuing role for prevention in high risk groups. Sexually Transmitted Infections 2006; 82: 368-371
6. Rakwar J, Lavreys L, Thompson ML et al. Cofactors for the acquisition of HIV-1 among heterosexual men: prospective cohort study of trucking company workers in Kenya. AIDS 1999; 13:607-614
7. Reid SR. Injection drug use, Unsafe medical injections and HIV in Africa: a systematic review. Harm Reduction Journal 6 (2009) 24. http://doi.org/10.1186/1477-7517-6-24
8. Heckathorn D. Respondent-driven sampling: a new approach to the study of hidden populations. Social problems 1992; 44:174-199
9. Li Y, Detels R, Lin P, et al. Difference in risk behaviours and STD prevalence between street-based and establishment based FSW in Guangdong Province, China. AIDS and Behaviour 2012; 16:943-951
10. Cowan F M, Mtetwa S, Davey C, Fearon E et al. Engagement with HIV prevention treatment and care among female sex workers in Zimbabwe. A respondent driven sampling survey. PLoS One 2013; 15, 8(10)
11. Skinner C, Mason B. Weighting in the regression analysis of survey data with a cross-national application. Personal.lse.ac.uk/skinnecj/CanJ2012.pdf. Accessed 25/05/2015. 13.16hrs.
12. Taruberekera N, Mishra V, Gonese E and Mugurungi O. Risk-taking behaviours of HIV- positive adults in Zimbabwe: Opportunities for prevention

with the positives. Zimbabwe working papers. 2010; No. 3. Calverton, Maryland, USA: ICF Macro.

13. Bwayo J J, Omari A M, Mutere A N et al. Long distance truck-drivers: 1.Prevalence of sexually transmitted diseases (STDs). East African Journal of Medicine. 1991 Jun; 68(6):425-429
14. Coughlan E, Mindel A, Estcourt CS. Male clients of female commercial sex workers:HIV, STDs and risk behaviour. International Journal of STD, AIDS 2001; 12 665-669
15. HIV Risk among adult sex workers in the United States. Centre for Disease Control and Prevention. September 2013. www.cdc.gov/hiv/pdf/library_factsheet_HIV_among_sex_workers.pdf. Accessed 25/05/2015. 10.18hrs.
16. Chen X S, Yin Y P, Gong X D, Liang G J et al. Prevalence of sexually transmitted infections among long distance truck drivers in Tongling, China. International Journal of STIs, AIDS. 2006 May; 17(5): 304-308.
17. Bwayo JJ, Plummer F, Omari M, et al. Human Immunodeficiency Virus Infection in Long Distance Truck Drivers in East Africa. Arch Intern Medicine 1994; 154(June 27): 1391-1396.
18. Gibney L, Saquib N, Metzger J. Behavioral risk factors for STD/HIV transmission in Bangladesh's trucking industry. Social Science and Medicine 2003; 56:1411-1424
19. Sagguti N, Schensul SL, Singh R. Alcohol use, sexual risk behaviour and STIs among married men in Mumbai, India. AIDS and Behaviour 2010; 14: 38-45
20. Sivaram S, Latkin CA, Solomon S, Celentano DD. HIV prevention in India: focus on men, alcohol use and social networks. Harvard Health Policy Review 2006; 7(2):125-134
21. Chatirvedi S, Singh Z, Banerjee A, et al. Sexual behaviour among long distance truck drivers. Indian Journal of Community Medicine 2006; 31(3): 153-156
22. Johnson D R. Using weights in the analysis of survey data. Population research institute, Department of Sociology, Pennsylvania state university. 2008.

23. Skinner C, Mason B. Weighting in the regression analysis of survey data with a cross-national application. Personal.lse.ac.uk/skinnecj/CanJ2012.pdf. Accessed 25/05/2015. 13.16hrs.
24. Fuller W A (2009). Sampling statistics,Wiley, Hoboken.
25. Hosmer DW and Lemeshow S. Applied Logistic Regression. Second edition. 2000. John Wiley & Sons, Inc, Hoboken, New Jersey
26. Agresti A. An Introduction to Categorical Data Analysis. Second Edition.2007. John Wiley & Sons, Inc, Hoboken, New Jersey
27. StataCorp. 2009. Stata Release 11. Statistical Software. College Station, TX: StataCorp LP.
28. Baker K . An analysis of survey data to determine significant risk factors associated with adolescent marijuana use through utilisation of sample weighting methods. Master's Thesis, University of Pittsburgh. 2015
29. Lee I. Searching for appropriate models for survey data analysis. www.sagepub.com/upm-data/6428_Chapter_6_Lee_(Analyzing)_I_PDF_7.pdf Accessed 01.06.2015. 19.03hrs.
30. Kish L. Weighting: Why, when, and how? A survey for surveys. Proceedings of the section on Survey Research methods. American Statistical Association. 1990: 121-130
31. Poi B P. From the help desk: Some bootstrapping techniques. The Stata Journal 2004; 4(3):312-328
32. Burns Statistics. The statistical bootstrap and other resampling methods. www.burns-stat.com/documents/tutorials/the-statistical-bootstrap-and-other-resampling-methods-2/ Accessed 02/06/2015; 1622hrs.
33. Survey data analysis. Introduction to survey commands. The survey data reference manual. www.stata.com/features/survey-data/svy-survey.pdf#page=2 Accessed 02/06/2015 1633hrs
34. Archer JK, Lemeshow S. Goodness-of-fit test for a logistic regression model fitted using survey sample data. The Stata Journal; 2006: 6(1); 97-105
35. Approaches to the analysis of Survey Data. Statistical Services Centre; The University of Reading, UK. Biometrics Advisory and Support Service to DFID. 2001.

36. The theory of weighting in the analysis of survey data. National Centre for Research Methods. www.restore.ac.uk/PEAS/theoryweightingtxt.php. Accessed 15/08/2015 16.30 hrs.
37. Pfefferman D. The role of sampling weights when modelling survey data. International statistical review/ Revue Internationale de Statisque; 61(2), 1993:317-337.
38. Garland S M, Tabrizi S N. Diagnosis of sexually transmitted infections (STI) using self collected non-invasive specimens. Sex Health. 2004; 1(2):121-126.
39. WHO. Essential Medicines and Health Products Information Portal. A World Health Organisation Resource. Guidelines for the Management of Sexually Transmitted Infections. 2001. Apps.who.int/medicinedocs/en/d/jh2942e/2.4.html. Accessed 25/05/2015. 11.37hrs.
40. Survey data analysis in STATA. UCLA: Statistical Consulting Group. http://www.ats.ucla.edu/stat/Statistical_Computing_Seminars_Survey_Data_Analysis_with_Stata.html. Accessed 27/07/2015; 1600hrs
41. Setting seed in STATA. www.stata.com/manuals/rsetseed.pdf Accessed 27/07/2015; 1630hrs.
42. Chromy JR, Abeyasekera S. Statistical analysis of survey data. Chapter 19. Household surveys in developing and transition countries: Design and Implementation analysis. Research Triangle Park, North Carolina USA.2005
43. Suchindran CM. Sampling Weights and Regression Analysis Lecture Notes. Fellow Carolina Population Centre. 2013
44. Introduction to Survey commands. STATA Survey Data Reference Manual Release 13. A Stata Press Publication. StataCorp LP. College Station, Texas. 2013
45. Iannacchione VG, Milne JG and Folsom RE. Response probability weight adjustment using Logistic Regression. Research Triangle Institute. Proceedings of the American Statistical Association, Section on Survey Research Methods. 1992. 637-642
46. King G and Zeng L. Logistic Regression in Rare events Data. Society for Political Methodology. 2001.

47. Carlson BL. Software for statistical analysis of sample survey data. Design of experiments and sample surveys. Encyclopaedia of Biostatistics 1998. John Wiley and Sons Inc.
48. Delany – Moretlwe S, Bello B, Kinros P et al. HIV prevalence and risk in long distance truck drivers in South Africa: A national cross-sectional survey. International Journal of STD and AIDS. 2014 May; 25(6): 428-38
49. D Wilson, R Marima, N Dube et al. Corridors of Hope in Southern Africa: HIV prevention needs and opportunities in four border towns. November 1999. Family health International.
50. Braustein SL, Ingabire CM, Kestekyn E, et al. High HIV immunodeficiency virus incidence in a cohort of Rwandan female sex workers. Sexually Transmitted Diseases 2011; 38:1-10
51. Naing NN. Determination of sample size. The Malaysian Journal of Medical Sciences: MJMS, 10(2), 84-86.
52. Dias S, Gama A, Fuertes R, et al. risk taking behaviours and HIV infection among sex workers in Portugal: results from a cross sectional survey. Sexually Transmitted Infections 2014; 0:1-7.

ANNEXE 1

Taylor series linearization

The Taylor series (linearization) method is the most commonly used method to estimate the covariance matrix of the regression coefficients for complex survey data in statistical software[44]. It is used with variance estimation for statistics that are vastly more complex than mere additions of sample values. Two factors that complicate variance estimation are complex sample design features and non-linearity of many common statistical estimators from complex sample surveys. Complex design features include stratification, clustering, multi-stage sampling, unequal probability sampling and without replacement sampling[42]. Non-linear statistical estimators for complex surveys include means, proportions and regression coefficients. The precision of these estimators is evaluated by their variances, hence the use of Taylor linearization helps in improving the precision since it take into account the sampling design to calculate the variances.

In statistics, methods that are used to estimate population parameters and their associated variances are usually based on assumptions about the characteristics and underlying distribution of the observations[47]. One of the assumptions is that the observations were selected independently and that each observation had the same probability of being selected. This assumption is violated however, when surveys are conducted. It is imperative therefore to account for this violation during analysis stage so as to get the correct variance associated with an estimate.

Since most estimates of sample surveys are nonlinear, the Taylor series expansion linearize such estimates with an assumption that all higher order terms are of negligible size[42], leaving only the first order (linear) portion of the expanded estimate. A standard formula for the mean square error of a linear estimate can then be applied to the linearized version to approximate the variance of the estimate. It has been shown that this approximation works well to the extent that the assumption regarding the higher order terms is correct.

ANNEXE 2

Diagnostic plots for unweighted models

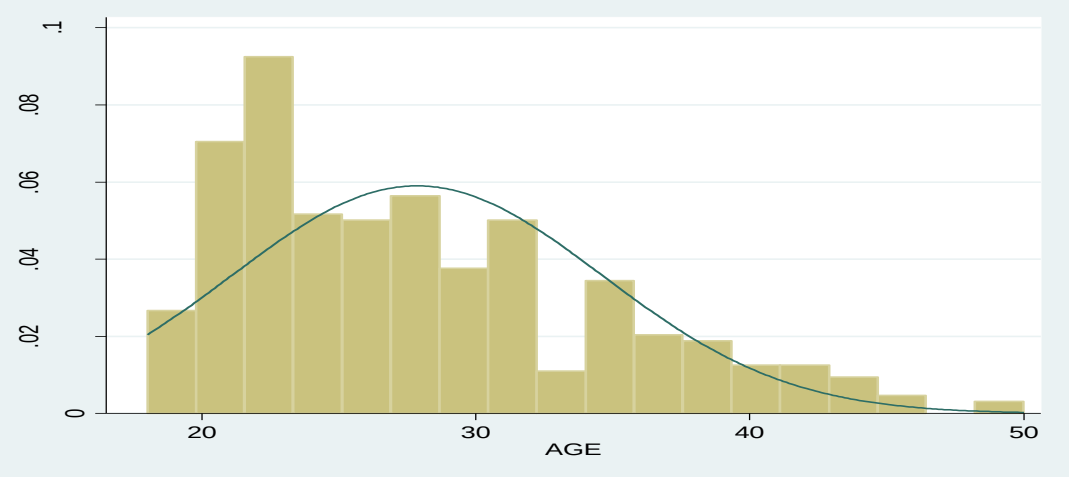
Sex workers

The final model for sex workers in the unweighted model is:

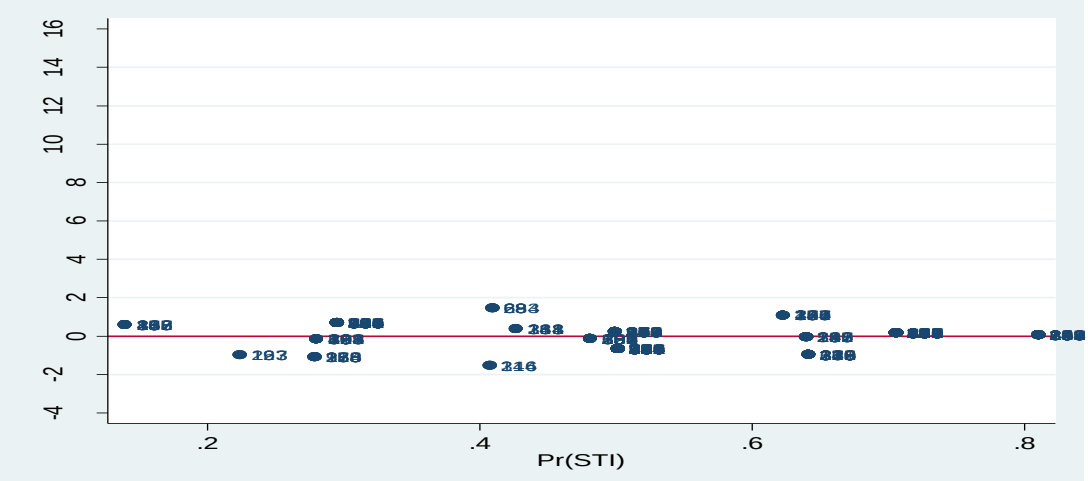
Logistic (STI) = 0.11 + 2.40 ALCOHOL + 1.78SEXDRUGS + 0.39USECON + 2.38PAIDEXTRA

(Where: *ALCOHOL is alcohol use; SEXDRUGS is used drugs to enhance sexual pleasure; USECON is used condom during the first sexual encounter; PAIDEXTRA is being paid extra not to use a condom*)

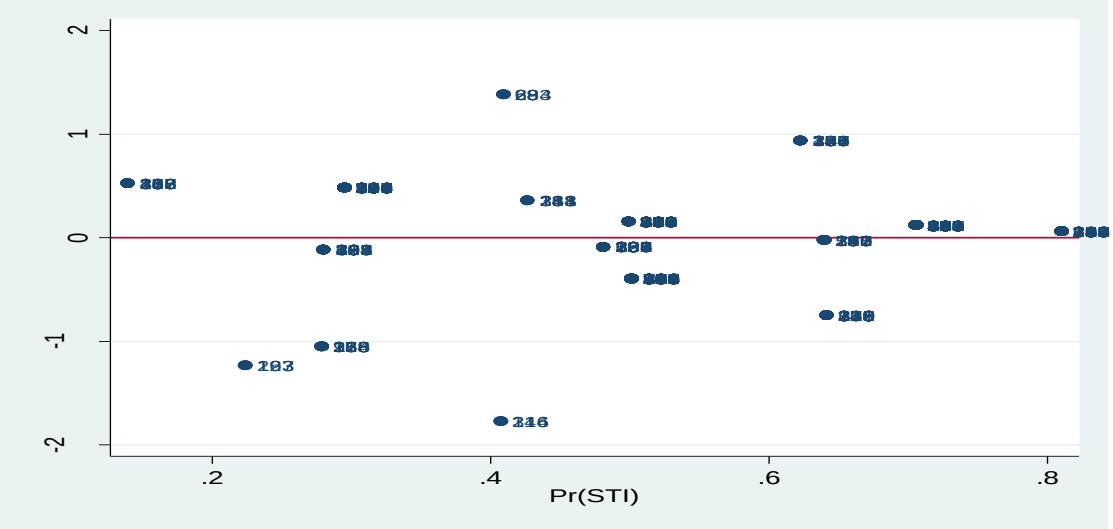
Histogram plot for AGE



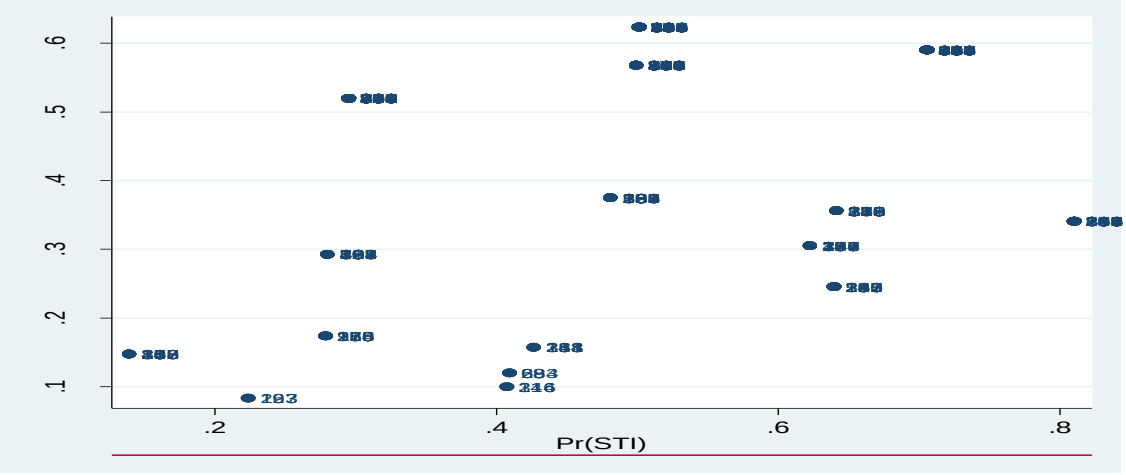
Scatter plot of standardized residuals against predicted probabilities



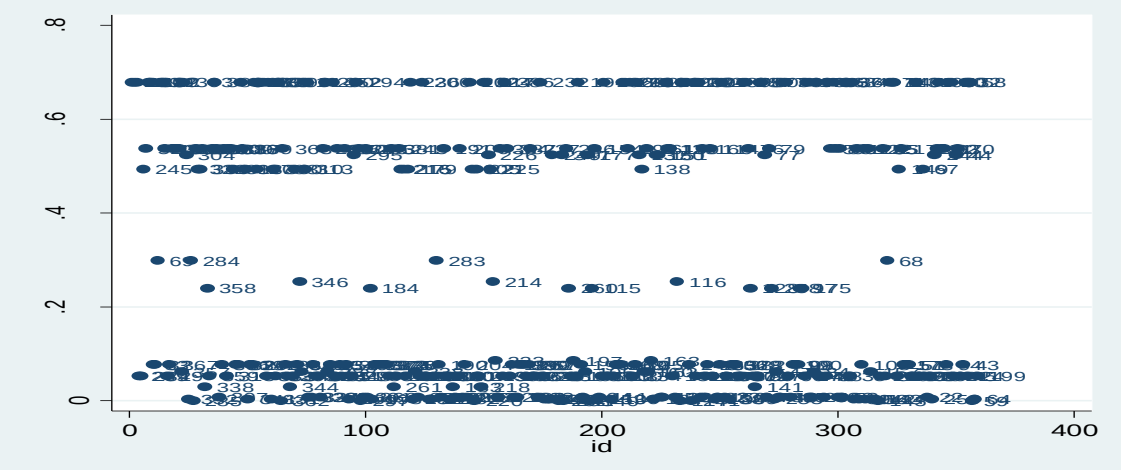
Scatter plot of deviance residuals against predicted probabilities



Scatter plot of leverage against predicted probabilities



Scatter plot of Pregibon's dbeta against each individual observation



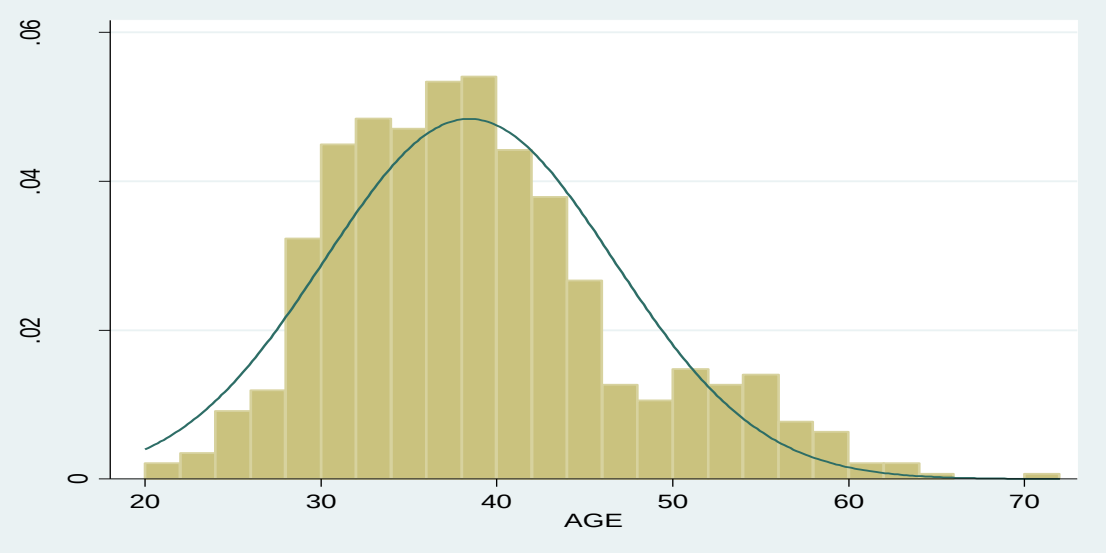
Long distance truck drivers

The final model for long distance truck drivers in the unweighted model was:

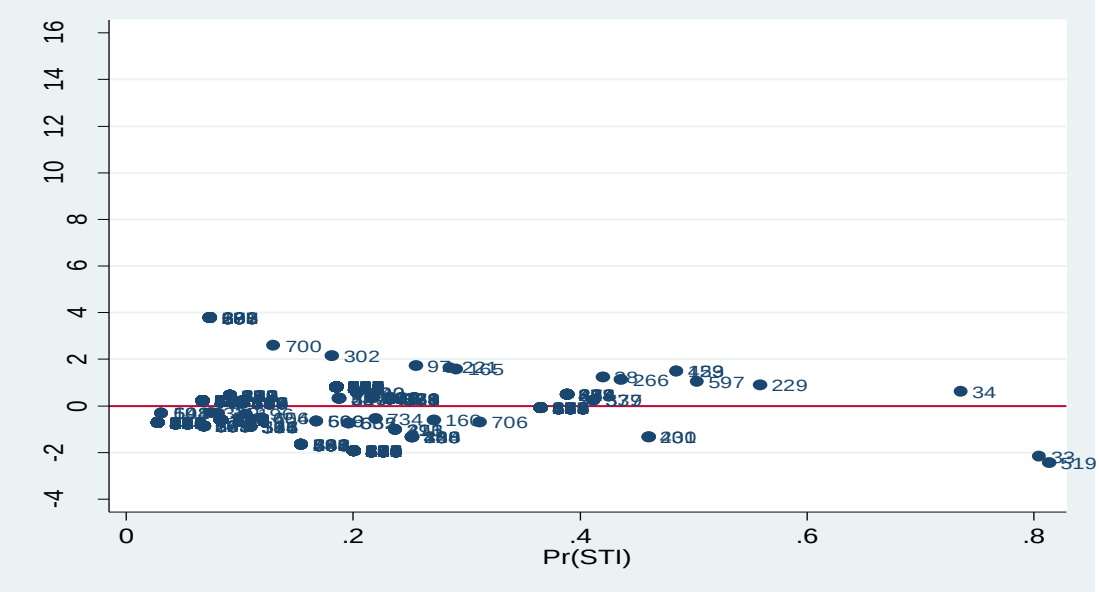
Logistic (STI) = 0.001 + 2.49 AFHM + 2.53 ETD + 1.10 NRFP + 3.15 USECON.

(Where AFHM is being away from home for more than a month; ETD is ever taken drugs; NRFP is number of non regular female sexual partners and USECON is used a male condom within the last 12 months during sex)

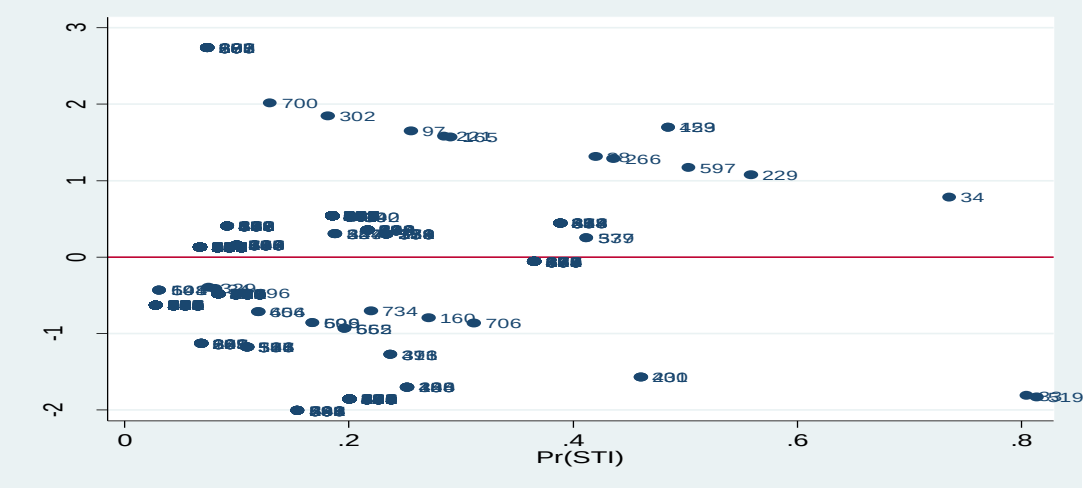
Histogram plot for AGE



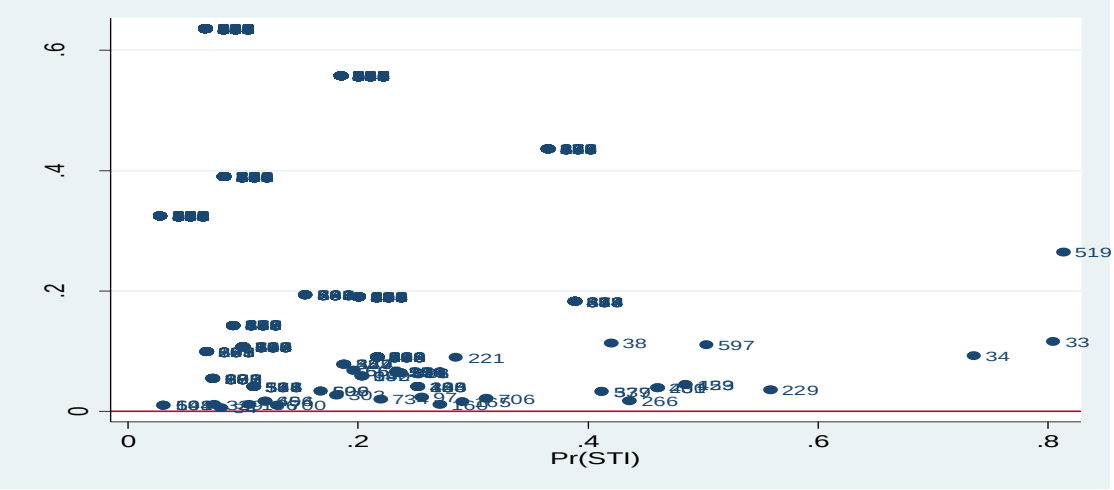
Scatter plot of standardized residuals against predicted probabilities



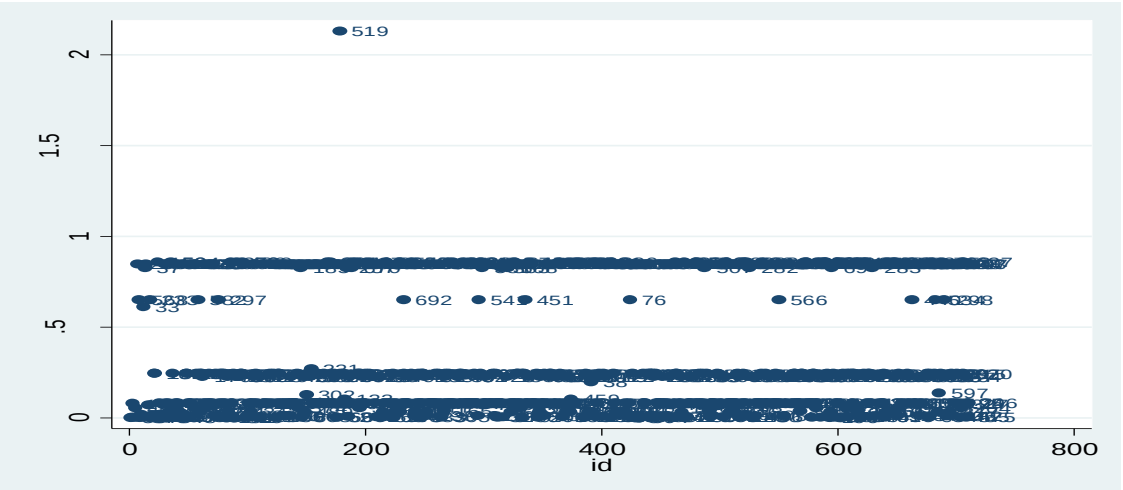
Scatter plot of deviance residuals against predicted probabilities



Scatter plot of leverage against predicted probabilities



Scatter plot of Pregibon's dbeta against each individual observation



ANNEXE 3

Variables selected for secondary data analysis

Data dictionary for SW

	Variable description	Variable code	Variable format	Variable name
Outcome variable	Sexually transmitted infections	Ever experienced signs and symptoms of an STI in the preceding 12 months 0 = no 1= yes	Binary	STI
Study variables	Age	Age (in years) at last birthday	Numeric	AGE
	Education level	Highest level of education attained 1 = at least primary 2 = junior school 3 = senior school and above	Nominal	EDULEVEL
	Religion	Which religion do you belong to? 1 = no religion 2 = traditional 3 = roman catholic 4 = Christian, non catholic 5 = Islam 6 = no response	Nominal	RELIGION
	Ethnicity	Which ethnic	Nominal	ETHNICITY

	group do you belong to? 1 = Ndebele 2 = shone 3 = venda 4 = other		
Employment status	Are you currently employed 1 = not employed 2 = part time employed 3 = full time employed	Nominal	EMPLOYED
Marital status	Current marital status 1 = single 2 = living together but not married 3 = married 4 = divorced/separated 5 = widowed	Nominal	MARITAL
Alcohol intake	Ever had drinks containing alcohol 1 = no 2 = yes	Binary	ALCOHOL
Age at first drink	Age when you started drinking alcohol	Numeric	AGEDRINK
Frequency of drinking	Frequency of drinking in the past 4 weeks	Nominal	ALCFREQ

	1 = never in the last 4 weeks 2 = less than once a week 3 = at least once a week 4 = everyday		
Drugs use	Ever taken drugs other than for the purpose of medical treatment 1 = no 2 = yes	Binary	DRUGS
Recent drug use	Used drugs in the past 12 months 1 = no 2 = yes	Binary	RDU
Cocaine use	Used cocaine in the past 12 months 1 = no 2 = yes	Binary	COCAINE
Cannabis use	Used cannabis in the past 12 months 1 = no 2 = yes	Binary	CANNABIS
Ecstasy use	Used ecstasy in the past 12 months 1 = no 2 = yes	Binary	ECSTASY
Other drugs used	Other drugs used in the past 12 months 1 = no	Binary	OTHER

	2 = yes		
Used drugs for sex	Ever taken drugs to increase sexual pleasure 1 = no 2 = yes	Binary	SEXDRUGS
Used cocaine	Used cocaine to increase sexual pleasure 1 = no 2 = yes	Binary	SEXCOC
Used Viagra	Used Viagra to increase sexual pleasure 1 = no 2 = yes	Binary	SEXVIAGRA
Used other drugs	Used other drugs to increase sexual pleasure 1 = no 2 = no	Binary	SEXOTHER
Age at first sex	Age first had sexual intercourse	Numeric	AFS
Condom use	Was condom used during the first sexual encounter 1 = no 2 = yes	Binary	USECON
First received money for sex	Age first received money for sex	Numeric	ARMFS
Condom use frequency	Frequency of condom use in the past 12 months	Nominal	FCONUSE

	1 = everytime 2 = almost everytime 3 = sometime 4 = never		
Suggesting condom use	Frequency of suggesting condom use in the past 12 months 1 = always 2 = sometimes 3 = never	Nominal	FSCONUSE
Condom use for anal intercourse	Frequency of using condoms if ever had anal intercourse with a client in the past 12 months 1 = no anal 2 = everytime 3 = almost everytime 4 = sometime 5 = never	Nominal	ANALCON
Suggesting condom use for anal intercourse	Frequency of suggesting condom use for anal intercourse in the past 12 months 1 = no anal 2 = always 3 = sometimes 4 = never	Nominal	FSANALCON

Paying clients	Number of paying clients in the last 7 days	Numeric	PAYING
Non paying clients	Number of non paying clients in the last 7 days	Numeric	NONPAYING
Condom use with paying clients	Frequency of condom use with paying clients in the last 7 days 1 = no sex past 7 days 2 = everytime 3 = almost everytime 4 = sometimes 5 = never	Nominal	PFCONUSE
Condom use for anal intercourse	Frequency of condom use for anal intercourse with paying client in the last 7 days 1 = no anal 2 = everytime 3 = almost everytime 4 = sometimes 5 = never	Nominal	PANALCON
Number of clients	Number of clients on the last day worked	Numeric	CLIENTS
Handshake	Had handshake sex with paying	Binary	PHSHAKE

	client the last time had se 1 = no 2 = yes		
Oral	Had oral sex (sucked client's penis) with paying client last time had sex 1 = no 2 = yes	Binary	PORAL
Oral by client	Had oral sex (client licked vagina) the last time had sex 1 = no 2 = yes	Binary	PORALBC
Intercrural	Had intercrural (thigh) sex with paying client last time had sex 1 = no 2 = yes	Binary	PINTERC
Vaginal	Had vaginal sex with paying client last time had sex 1 = no 2 = yes	Binary	PVAGINA
Condom use with paying client	Used a condom for vaginal sex the last time with paying client	Nominal	PCONUSE

	1 = no vaginal sex 2 = yes 3 = no		
Suggesting condom use	Who suggested condom use on the last time with paying client for vaginal sex 1 = myself 2 = client 3 = joint decision	Nominal	PCONSUG
Paid extra for no condom	Didn't use a condom with paying client because client had paid extra for no condom 1 = no 2 = yes	Binary	PAIDEXTRA
Regular non paying partners	Number of regular non paying partners	Numeric	NPR
Non regular non-paying partners	Number of non-regular non-paying partners	Numeric	NPNR
Condom use with non-paying partners	Frequency of condom use with non-paying partners in the past 12 months 1 = no non-paying partner 2 = everytime	Nominal	NPCONUSE

	3 = almost everytime 4 = sometimes 5 = never
Condom use with non-paying	Was a condom used for vaginal intercourse with a non-paying partner the last time had sex 1 = no 2 = yes

Data dictionary for LDTD

	Variable description	Variable code	Variable format	Variable name
Outcome variable	Sexually transmitted infections	Ever experienced signs and symptoms of an STI in the preceding 12 months 0 = no 1= yes	Binary	STI
Study variables	Age	Age (in years) at last birthday	Numeric	AGE
	Education level	Highest level of education attained 1 = at least primary 2 = junior school 3 = senior school 4 = tertiary	Nominal	EDULEVEL
	Religion	Which religion do you belong to? 1 = no religion 2 = traditional 3 = roman catholic 4 = Christian (non catholic)	Nominal	RELIGION

	5 = Islam 6 = other		
Ethnicity	Which ethnic group do you belong to? 1 = Bemba 2 = Himba 3 = Ndebele 4 = Shona 5 = Swazi 6 = Venda 7 = Xhosa 8 = Zulu 9 = other	Nominal	ETHNICITY
Marital status	Marital status 1 = single 2 = living together but not married 3 = married 4 = divorced/ separated 5 = widowed	Nominal	MARITAL
Length of service	Length of service as a truck driver in years	Numeric	LOSYRS
Length of service	Length of service as a truck driver in months	Numeric	LOSMNTS

Away from home	Ever been away from home for more than one month in the past 12 months 1 = no 2 = yes	Binary	AFHM
Length away from home	Length of stay away from home on the last occasion away	Numeric	LAFHM
Alcohol intake	Ever had drinks containing alcohol 1 = no 2 = yes	Binary	ALCOHOL
Age at first drink	Age when alcohol drinking started	Numeric	AFD
Drinking frequency	Frequency of alcohol intake in the past 4 weeks 1 = never drink alcohol 2 = never in the last 4 weeks	Nominal	ALCFREQ

	3 = less than once a week 4 = at least once a week 5 = everyday		
Ever had drugs	Ever taken any drugs other than for medical reasons 1 = no 2 = yes	Binary	ETD
Recent drug use	Took drugs in the past 12 months other than for medical reasons 1 = no 2 = yes	Binary	DRUGS
Cannabis use	Used cannabis in the past 12 months 1 = no 2 = yes	Binary	CANNABIS
Cocaine use	Used cocaine in the past 12 months 1 = no 2 = yes	Binary	COCAINE
Ecstasy use	Used ecstasy in the past 12	Binary	ECSTASY

	months 1 = no 2 = yes		
Amphetamine use	Used amphetamine in the past 12 months 1 = no 2 = yes	Binary	AMPH
Opium use	Used opium in the past 12 months 1 = no 2 = yes	Binary	OPIUM
Hashish use	Used hashish in the past 12 months 1 = no 2 = yes	Binary	HASH
Crystal meth use	Used crystal meth in the past 12 months 1 = no 2 = yes	Binary	CRYSTAL
Heroin use	Used heroin in the past 12 months 1 = no 2 = yes	Binary	HEROIN
Other drugs used	Used other drugs in the past 12	Binary	Other

	months 1 = no 2 = yes		
Drugs for sex	Ever taken drugs to increase sexual pleasure 1 = no 2= yes	Binary	DRUGSEX
Cocaine	Used cocaine to increase sexual pleasure 1 = no 2 = yes	Binary	SEXCOC
Ecstasy	Used ecstasy to increase sexual pleasure 1 = no 2 = yes	Binary	SEXECS
Amphetamines	Used amphetamines to increase sexual pleasure 1 = no 2 = yes	Binary	SEXAMPH
Opium	Used opium to increase sexual pleasure	Binary	SEXOP

	1 = no 2 = yes		
Hashish	Used hashish to increase sexual pleasure 1 = no 2 = yes	Binary	SEXHASH
Crystal meth	Used crystal meth to increase sexual pleasure 1 = no 2 = yes	Binary	SEXCRYST
Heroin	Used heroin to increase sexual pleasure 1 = no 2 = yes	Binary	SEXHEROIN
Viagra	Used Viagra to increase sexual pleasure 1 = no 2 = yes	Binary	SEXVIAG
Other drugs	Used other drugs to increase sexual pleasure 1 = no	Binary	SEXOTHER

	2 = yes		
Injecting drugs	Ever injected drugs other than as part of prescribed medical treatment in the past 12 months 1 = no 2 = yes	Binary	IDU
Had sex	Ever had sexual intercourse 1 = no 2 = yes	Binary	HADSEX
Age at first sex	Age at the first sexual encounter	Numeric	AFS
Condom use	Condom use at first sexual encounter 1 = no 2 = yes	Binary	CONUSE
Recent sexual activity	Had sexual intercourse in the past 12 months 1 = no 2 = yes	Binary	SEXHAD
Sex with female	Ever had sexual	Binary	SEXFEM

	intercourse with a female 1 = no 2 = yes		
Regular partners	Number of regular female partners in the past 12 months	Numeric	RFP
Non regular partners	Number of non-regular female partners in the past 12 months	Numeric	NRFP
Commercial partners	Number commercial female partners in the past 12 months	Numeric	CFP
Condom use with regular partner	Frequency of condom use with regular partner in the past 12 months 1 = everytime 2 = almost everytime 3 = sometime 4 = never 5 = no regular	Nominal	RPCONUSE

	partner		
Condom use on the last time	Used a condom with regular partner for vaginal intercourse the last time 1 = no 2 = yes	Binary	CONUSELT
Condom use for anal sex	Frequency of condom use with regular partner for anal sex in the past 12 months 1 = no anal 2 = everytime 3 = sometimes 4 = never	Nominal	FRPANAL
Condom use with non-regular partner	Frequency of condom use for vaginal intercourse with non-regular female partner 1 = everytime 2 = almost everytime 3 = sometimes 4 = never 5 = don’t know	Nominal	NRPCONUSE

	6 = no non-regular partner 7 = no vaginal intercourse with non-regular partner		
Condom use on the last time	Used a condom for vaginal intercourse with a non regular partner on the last time 1 = no 2 = yes	Binary	NRPLT
Condom use with commercial partners	Frequency of condom use with commercial partners in the past 12 months 1 = everytime 2 = almost everytime 3 = sometimes 4 = never 5 = don’t know 6 = no commercial partners 7 = no vaginal	Nominal	CPCONUSE

	intercourse		
Condom use last time	Used condom for vaginal intercourse with a commercial partner on the last time 1 = no 2 = yes	Binary	CPLT
Condom use for anal intercourse	Frequency of condom use for anal intercourse in the past 12 months 1 = everytime 2 = almost everytime 3 = never 4 = no response 5 = no anal	Nominal	CPANAL
Circumcised	Have you been circumcised 1 = no 2 = yes	Binary	CIRCUM
Condom use past 12 months	Ever used a male condom within the past 12 months during sex	Binary	USECON

	1 = no 2 = yes		
Carry condoms	Always carry condoms when travelling 1 = no 2 = yes	Binary	CARCON

ANNEXE 4

Do files

1. **Sex workers unweighted analysis**

```
*******SW Unweighted Analysis****

cd "D:\\2014 WHO STUDY"

set more off

cap log close

log using SWUnweighted.log, replace


use ZimbabweSW17Jul_2.dta, clear

**********************************

********Data recoding******

gen AGE= A1

gen EDULEVEL= 1 if A2==1| A2==2

replace EDULEVEL=2 if A2==3

replace EDULEVEL=3 if A2==4| A2==5

gen RELIGION= A4

gen ETHNICITY=1 if A5==6

replace ETHNICITY=2 if A5==8

replace ETHNICITY=3 if A5==10

replace ETHNICITY=4 if A5==14

gen EMPLOYED=1 if A7==1| A7==4

replace EMPLOYED=2 if A7==2

replace EMPLOYED=3 if A7==3

gen MARITAL= A8

gen ALCOHOL=1 if B1==2

replace ALCOHOL=2 if B1==1
```

```
gen AGEDRINK= B2
gen ALCFREQ=1 if B3==1| B3==2
replace ALCFREQ=2 if B3==3
replace ALCFREQ=3 if B3==4
replace ALCFREQ=4 if B3==5
gen DRUGS=1 if B4==1
replace DRUGS=2 if B4==2
gen RDU=1 if B5==2
replace RDU=2 if B5==1
gen CANNABIS=1 if B6_Cannabis==2
replace CANNABIS=2 if B6_Cannabis==1
gen COCAINE=1 if B6_Cocaine==2
replace COCAINE=2 if B6_Cocaine==1
gen ECSTASY=1 if B6_Ecstasy==2
replace ECSTASY=2 if B6_Ecstasy==1
gen OTHER=1 if B6_Other1==2
replace OTHER=2 if B6_Other1==1
gen SEXDRUGS=1 if B7==2
replace SEXDRUGS=2 if B7==1
gen SEXCOC=1 if B8_Cocaine==2
replace SEXCOC=2 if B8_Cocaine==1
gen SEXVIAGRA=1 if B8_Viagra==2
replace SEXVIAGRA=2 if B8_Viagra==1
gen SEXOTHER=1 if B8_Other1==2
replace SEXOTHER=2 if B8_Other1==1
gen AFS= C2
gen USECON=1 if C4==2
```

```
replace USECON=2 if C4==1

gen ARMFS= C6

gen FCONUSE= C9

gen FSCONUSE= C10

gen ANALCON=1 if C11==7

replace ANALCON=2 if C11==1

replace ANALCON=3 if C11==2

replace ANALCON=4 if C11==3

replace ANALCON=5 if C11==4

gen FSANALCON=1 if  C12==6

replace FSANALCON=2 if C12==1

replace FSANALCON=3 if C12==2

replace FSANALCON=4 if C12==3

gen PAYING= C15_PAYING_CLIENTS

gen NONPAYING= C15_NON_PAYING_PARTNERS

gen PFCONUSE= C16

gen PANALCON= C17

gen CLIENTS= C18

gen PHSHAKE = 1 if C21_r_HANDSHAKE___YOU_MA==2

replace PHSHAKE=2 if C21_r_HANDSHAKE___YOU_MA==1

gen PORAL = 1 if C21_ORAL_SEX__YOU_SUCKED==2

replace PORAL = 2 if C21_ORAL_SEX__YOU_SUCKED==1

gen PORALBC = 1 if C21_ORAL_SEX__CLIENT_LIC==2

replace PORALBC = 2 if C21_ORAL_SEX__CLIENT_LIC==1

gen PINTERC = 1 if C21_INTERCRURAL_SEX__THI==2

replace PINTERC = 2 if C21_INTERCRURAL_SEX__THI==1

gen PVAGINA = 1 if C21_VAGINAL_INTERCOURSE==2
```

```
replace PVAGINA = 2 if C21_VAGINAL_INTERCOURSE==1
gen PANAL = 1 if C21_ANAL_INTERCOURSE==2
replace PANAL = 2 if C21_ANAL_INTERCOURSE==1
gen PCONUSE = C22
gen PCONSUG=1 if C23==1
replace PCONSUG=2 if C23==2
replace PCONSUG = 3 if C23==3
gen PAIDEXTRA=1 if C28==2
replace PAIDEXTRA=2 if C28==1
gen NPR = C29_NUMBER_OF_REGULAR_PA
gen NPNR = C29_NUMBER_OF_NON_REGULA
gen NPCONUSE=1 if C30==1
replace NPCONUSE=2 if C30==3
replace NPCONUSE =3 if C30==4
replace NPCONUSE = 4 if C30==5
replace NPCONUSE = 5 if C30==6
gen NPCON =1 if C31==3
replace NPCON=2 if C31==2
*******************************************
**Demographic variables*
hist AGE, norm
sum AGE
univar AGE
tab EDULEVEL, miss
tab RELIGION, miss
tab ETHNICITY, miss
tab EMPLOYED, miss
```

```
tab MARITAL, miss
**********************************
***Univariate analysis****
logistic STI ALCOHOL
logistic STI AGEDRINK
xi: logistic STI i.ALCFREQ
logistic STI DRUGS
logistic STI RDU
logistic STI CANNABIS
logistic STI COCAINE
logistic STI ECSTASY
logistic STI OTHER
logistic STI SEXDRUGS
logistic STI SEXCOC
logistic STI SEXVIAGRA
logistic STI SEXOTHER
logistic STI AFS
logistic STI USECON
logistic STI ARMFS
xi: logistic STI i.FCONUSE
xi: logistic STI i.FSCONUSE
xi: logistic STI i.ANALCON
xi: logistic STI i.FSANALCON
logistic STI PAYING
logistic STI NONPAYING
xi: logistic STI i.PFCONUSE
xi: logistic STI i.PANALCON
```

```
logistic STI CLIENTS
logistic STI PHSHAKE
logistic STI PORAL
logistic STI PORALBC
logistic STI PINTERC
logistic STI PVAGINA
logistic STI PANAL
xi: logistic STI i.PCONUSE
xi: logistic STI i.PCONSUG
logistic STI PAIDEXTRA
logistic STI NPR
logistic STI NPNR
xi: logistic STI i.NPCONUSE
logistic STI NPCON
logistic STI AGE
xi: logistic STI i.EDULEVEL
xi: logistic STI i.MARITAL
xi: logistic STI i.ETHNICITY
xi: logistic STI i.RELIGION
xi: logistic STI i.EMPLOYED
**********************************************
*****Multivariate analysis****
logistic STI ALCOHOL
logistic STI ALCOHOL RDU
logistic STI ALCOHOL DRUGS
logistic STI ALCOHOL OTHER
logistic STI ALCOHOL SEXDRUGS
```

```
logistic STI ALCOHOL SEXDRUGS SEXVIAGRA
logistic STI ALCOHOL SEXDRUGS SEXOTHER
logistic STI ALCOHOL SEXDRUGS AFS
logistic STI ALCOHOL SEXDRUGS USECON
logistic STI ALCOHOL SEXDRUGS USECON PAYING
logistic STI ALCOHOL SEXDRUGS USECON CLIENTS
logistic STI ALCOHOL SEXDRUGS USECON PAIDEXTRA
logistic STI ALCOHOL SEXDRUGS USECON PAIDEXTRA NPCON
logistic STI ALCOHOL SEXDRUGS USECON PAIDEXTRA
**********************************************
*********Diagnostics &GOF**********
logistic STI ALCOHOL SEXDRUGS USECON PAIDEXTRA
estat classification
estat gof
linktest
lfit
lfit, group(10)table
fitstat
collin ALCOHOL SEXDRUGS USECON PAIDEXTRA
predict p
predict stdres, rstand
scatter stdres p,  mlabel(RespID) ylab(-4(2) 16) yline(0)
gen id= _n
scatter stdres id,  mlab(RespID) ylab(-4(2) 16) yline(0)
predict dv, dev
scatter dv p, mlab(RespID) yline(0)
scatter dv id, mlab(RespID)
```

```
predict hat, hat
scatter hat p, mlab( RespID) yline(0)
scatter hat id, mlab(RespID)
predict dbeta, dbeta
scatter dbeta id, mlab(RespID)
lroc
*****************************
******LR Test******
logistic STI
estimates store m1
logistic STI ALCOHOL SEXDRUGS USECON PAIDEXTRA
estimates store m2
lrtest m1 m2
test ALCOHOL SEXDRUGS USECON PAIDEXTRA
logistic STI ALCOHOL SEXDRUGS USECON PAIDEXTRA
coefplot, drop(_cons) xline(1) eform xtitle(Odds ratio)
*****************************
capture log close
```

2. Sex workers weighted analysis

```
*********SW Weighted Analysis*******
cd "D:\\WeightedLR"
set more off
cap log close
log using WeightedLRSW.log, replace

use ZimbabweSW17Jul_2.dta, clear
**********Setting weights*********
```

```
count
set seed 1003002849
sample 90
count
gen pw= 359/323
gen fpc= 359
svyset [pweight=pw], fpc(fpc)
******Demographics**********
svydes
svy: mean AGE
svy: tab EDULEVEL
svy: tab RELIGION
svy: tab ETHNICITY
svy: tab EMPLOYED
svy: tab MARITAL
**********Univariate analysis*****
svy: logistic STI ALCOHOL
svy: logistic STI AGEDRINK
xi: svy: logistic STI i.ALCFREQ
svy: logistic STI DRUGS
svy: logistic STI RDU
svy: logistic STI CANNABIS
svy: logistic STI COCAINE
svy: logistic STI ECSTASY
svy: logistic STI OTHER
svy: logistic STI SEXDRUGS
svy: logistic STI SEXCOC
```

svy: logistic STI SEXVIAGRA

svy: logistic STI SEXOTHER

svy: logistic STI AFS

svy: logistic STI USECON

svy: logistic STI ARMFS

xi:svy: logistic STI i.FCONUSE

xi:svy: logistic STI i.FSCONUSE

xi:svy: logistic STI i.ANALCON

xi:svy: logistic STI i.FSANALCON

svy: logistic STI PAYING

svy: logistic STI NONPAYING

xi:svy: logistic STI i.PFCONUSE

xi:svy: logistic STI i.PANALCON

svy: logistic STI CLIENTS

svy: logistic STI PHSHAKE

svy: logistic STI PORAL

svy: logistic STI PORALBC

svy: logistic STI PINTERC

svy: logistic STI PVAGINA

svy: logistic STI PANAL

xi:svy: logistic STI i.PCONUSE

xi:svy: logistic STI i.PCONSUG

svy: logistic STI PAIDEXTRA

svy: logistic STI NPR

svy: logistic STI NPNR

xi:svy: logistic STI i.NPCONUSE

svy: logistic STI NPCON

svy: logistic STI AGE

xi: svy: logistic STI i.EDULEVEL

xi: svy: logistic STI i.RELIGION

xi: svy: logistic STI i.ETHNICITY

xi: svy: logistic STI i.EMPLOYED

xi: svy: logistic STI i.MARITAL

****

*******Multivariate analysis**********

svy: logistic STI ALCOHOL

svy: logistic STI ALCOHOL AGEDRINK

svy: logistic STI ALCOHOL DRUGS

svy: logistic STI ALCOHOL DRUGS RDU

svy: logistic STI ALCOHOL RDU

svy: logistic STI ALCOHOL RDU CANNABIS

svy: logistic STI ALCOHOL RDU OTHER

svy: logistic STI ALCOHOL RDU SEXDRUGS

svy: logistic STI ALCOHOL RDU SEXDRUGS SEXVIAGRA

svy: logistic STI ALCOHOL RDU SEXDRUGS SEXOTHER

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS

xi: svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS i.ANALCON

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS PHSHAKE

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS PORALBC

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS PINTERC

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS PANAL

xi: svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS i.PCONSUG

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS PAIDEXTRA

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS PAIDEXTRA NPR

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS PAIDEXTRA NPNR

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS PAIDEXTRA NPNR NPCON

svy: logistic STI ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS PAIDEXTRA NPNR

****

**********Diagnostic &GOF tests**********

estat gof

svylogitgof

linktest

collin ALCOHOL RDU SEXDRUGS AFS USECON ARMFS PAYING CLIENTS PAIDEXTRA NPNR

****

****Dropping AFS ARMFS PAYING NPNR due to lack of fit*********

svy: logistic STI ALCOHOL RDU SEXDRUGS USECON PAIDEXTRA

estat gof

svylogitgof

```
linktest
test ALCOHOL RDU SEXDRUGS USECON PAIDEXTRA
collin ALCOHOL RDU SEXDRUGS USECON PAIDEXTRA
svy: logistic STI ALCOHOL RDU SEXDRUGS USECON PAIDEXTRA
coefplot, drop(_cons) xline(1) eform xtitle(Odds ratio)
**************************************************
capture log close
```

3. Long distance truck drivers unweighted analysis

```
*********LDTD Unweighted analysis****
cd "D:\\2014 WHO STUDY"
set more off
cap log close
log using UnweightedLDTD.log, replace

use ZLDTDfinal18Jul, clear
**************************************************
****Data recoding******
gen AGE= A1
gen EDULEVEL=1 if A2==1| A2==2
replace EDULEVEL=2 if A2==3
```

```
replace EDULEVEL=3 if A2==4

replace EDULEVEL=4 if A2==5| A2==6

gen RELIGION= A4

gen ETHNICITY= A5

gen MARITAL= A7

gen LOSYRS = A10_YEARS

gen LOSMNTS= A10_MONTHS

gen AFHM=1 if A11==2

replace AFHM=2 if A11==1

gen LAFHM= A12

gen ALCOHOL=1 if B1==2

replace ALCOHOL=2 if B1==1

gen AFD = B2

gen ALCFREQ = 1 if B3==1| B3==.

replace ALCFREQ = 2 if B3==2

replace ALCFREQ = 3 if B3==3

replace ALCFREQ= 4 if B3==4

replace ALCFREQ = 5 if B3==5

gen ETD = B4

gen DRUGS=1 if B5==2

replace DRUGS=2 if B5==1

gen CANNABIS=1 if B6_Cannabis==2

replace CANNABIS = 2 if B6_Cannabis==1

gen COCAINE=1 if B6_Cocaine==2

replace COCAINE = 2 if B6_Cocaine==1

gen ECSTASY = 1 if B6_Ecstasy==2

replace ECSTASY = 2 if B6_Ecstasy==1
```

```
gen AMPH = 1 if B6_Amphetamines==2
replace AMPH = 2 if B6_Amphetamines==1
gen OPIUM = 1 if B6_Opium==2
replace OPIUM = 2 if B6_Opium==1
gen HASH= 1 if B6_Hashish==2
replace HASH = 2 if B6_Hashish==1
gen CRYSTAL = 1 if B6_Crystal_Meth==2
replace CRYSTAL = 2 if B6_Crystal_Meth==1
gen HEROIN = 1 if B6_Heroin==2
replace HEROIN = 2 if B6_Heroin==1
gen OTHER = 1 if B6_Other1==2
replace OTHER = 2 if B6_Other1==1
gen DRUGSEX = 1 if B7==2
replace DRUGSEX = 2 if B7==1
gen SEXCOC = 1 if B8_Cocaine==2
replace SEXCOC = 2 if B8_Cocaine==1
gen SEXECS = 1 if B8_Ecstasy==2
replace SEXECS = 2 if B8_Ecstasy==1
gen SEXAMPH = 1 if B8_Amphetamines==2
replace SEXAMPH = 2 if B8_Amphetamines==1
gen SEXOP = 1 if B8_Opium==2
replace SEXOP = 2 if B8_Opium==1
gen SEXHASH = 1 if B8_Hashish==2
replace SEXHASH = 2 if B8_Hashish==1
gen SEXCRYST = 1 if B8_Crystal_Meth==2
replace SEXCRYST = 2 if B8_Crystal_Meth==1
gen SEXHEROIN=1 if B8_Heroin==2
```

```
replace SEXHEROIN=2 if B8_Heroin==1

gen SEXVIAG = 1 if B8_Viagra==2

replace SEXVIAG=2 if B8_Viagra==1

gen SEXOTHER = 1 if B8_Other1==2

replace SEXOTHER=2 if B8_Other1==1

gen IDU = 1 if B9==2

replace IDU = 2 if B9==1

gen HADSEX = 1 if C1==2

replace HADSEX = 2 if C1==1

gen AFS = C2

gen CONUSE = 1 if C4==2

replace CONUSE = 2 if C4==1

gen SEXHAD= 1 if C5==2

replace SEXHAD = 2 if C5==1

gen SEXFEM= 1 if C6==2

replace SEXFEM = 2 if C6==1

gen RFP = C7_REGULAR

gen NRFP = C7_NON_REGULAR

gen CFP = C7_COMMERCIAL

gen RPCONUSE = 1 if C8==1

replace RPCONUSE = 2 if C8==2

replace RPCONUSE = 3 if C8==3

replace RPCONUSE = 4 if C8==4

replace RPCONUSE = 5 if C8==5| C8==6| C8==.

gen CONUSELT = 1 if C9==2

replace CONUSELT = 2 if C9==1

gen RFPANAL = 1 if C11==7
```

```
replace RFPANAL = 2 if C11==1
replace RFPANAL = 3 if C11==3
replace RFPANAL = 4 if C11==4
gen NRPCONUSE = 1 if C14==1
replace NRPCONUSE = 2 if C14==2
replace NRPCONUSE = 3 if C14==3
replace NRPCONUSE = 4 if C14==4
replace NRPCONUSE = 5 if C14==5| C14==6| C14==.
replace NRPCONUSE = 6 if C14==7
replace NRPCONUSE = 7 if C14==8
gen NRPLT = 1 if C15==2
replace NRPLT = 2 if C15==1
gen CPCONUSE = 1 if C20==1
replace CPCONUSE = 2 if C20==2
replace CPCONUSE = 3 if C20==3
replace CPCONUSE = 4 if C20==4
replace CPCONUSE = 5 if C20==5| C20==.
replace CPCONUSE = 6 if C20==7
replace CPCONUSE = 7 if C20==8
gen CPLT = 1 if C21==2
replace CPLT = 2 if C21==1
gen CPANAL = 1 if C23==1
replace CPANAL = 2 if C23==2
replace CPANAL = 3 if C23==4
replace CPANAL = 4 if C23==6| C23==.
replace CPANAL = 5 if C23==7
gen CIRCUM = 1 if C43==2
```

```
replace CIRCUM = 2 if C43==1
gen USECON = 1 if D1==2
replace USECON = 2 if D1==1
gen CARCON = 1 if D4==2
replace CARCON = 2 if D4==1
*****************************************
*********Demographics**********************
hist AGE , norm
univar AGE
mean AGE
tab EDULEVEL, miss
tab RELIGION, miss
tab MARITAL, miss
tab ETHNICITY, miss
hist LOSYRS, norm
mean LOSYRS
univar LOSYRS
hist LOSMNTS, norm
univar LOSMNTS
mean LOSMNTS
************************************************
**********Univariate analysis***********
logistic STI LOSYRS
logistic STI LOSMNTS
logistic STI AFHM
logistic STI LAFHM
logistic STI ALCOHOL
```

```
logistic STI AFD
xi: logistic STI i.ALCFREQ
logistic STI ETD
logistic STI DRUGS
logistic STI CANNABIS
logistic STI COCAINE
logistic STI ECSTASY
logistic STI AMPH
logistic STI OPIUM
logistic STI HASH
logistic STI CRYSTAL
logistic STI HEROIN
logistic STI OTHER
logistic STI DRUGSEX
logistic STI SEXCOC
logistic STI SEXECS
logistic STI SEXAMPH
logistic STI SEXOP
logistic STI SEXHASH
logistic STI SEXCRYST
logistic STI SEXHEROIN
logistic STI SEXVIAG
logistic STI SEXOTHER
logistic STI IDU
logistic STI HADSEX
logistic STI AFS
logistic STI CONUSE
```

logistic STI SEXHAD

logistic STI SEXFEM

logistic STI RFP

logistic STI NRFP

logistic STI CFP

xi: logistic STI i.RPCONUSE

logistic STI CONUSELT

xi: logistic STI i.RFPANAL

xi: logistic STI i.NRPCONUSE

xi: logistic STI i.CPCONUSE

logistic STI CPLT

xi: logistic STI i.CPANAL

logistic STI CIRCUM

logistic STI USECON

logistic STI AGE

xi: logistic STI i.EDULEVEL

xi: logistic STI i.RELIGION

xi: logistic STI i.MARITAL

xi: logistic STI i.ETHNICITY

****

**********Multivariate analysis*********

logistic STI AFHM LAFHM

logistic STI AFHM ALCOHOL

logistic STI AFHM ALCOHOL AFD

logistic STI AFHM ALCOHOL ETD

logistic STI AFHM ETD

logistic STI AFHM ETD DRUGS

```
logistic STI AFHM ETD CANNABIS
logistic STI AFHM ETD COCAINE
logistic STI AFHM ETD DRUGSEX
logistic STI AFHM ETD IDU
logistic STI AFHM ETD AFS
logistic STI AFHM ETD RFP
logistic STI AFHM ETD NRFP
logistic STI AFHM ETD NRFP USECON
logistic STI AFHM ETD NRFP USECON CARCON
****************************************
*********Diagnostic& GOF**********
logistic STI AFHM ETD NRFP USECON
estat classification
estat gof
linktest
lfit
lfit, group(10)table
fitstat
collin AFHM ETD NRFP USECON
predict p
predict stdres, rstand
scatter stdres p,  mlabel(RespID) ylab(-4(2) 16) yline(0)
gen id= _n
scatter stdres id,  mlab(RespID) ylab(-4(2) 16) yline(0)
predict dv, dev
scatter dv p, mlab(RespID) yline(0)
scatter dv id, mlab(RespID)
```

```
predict hat, hat
scatter hat p, mlab( RespID) yline(0)
scatter hat id, mlab(RespID)
predict dbeta, dbeta
scatter dbeta id, mlab(RespID)
lroc
******LR test*******
logistic STI
estimates store m1
logistic STI AFHM ETD NRFP USECON
estimates store m2
lrtest m1 m2
test AFHM ETD NRFP USECON
logistic STI AFHM ETD NRFP USECON
coefplot, drop(_cons) xline(1) eform xtitle(Odds ratio)
*************************************
capture log close
```

4. Long distance truck drivers weighted analysis

```
******************LDTD Weighted Analysis******************
cd "D:\\WeightedLR"
set more off
cap log close
log using WeightedLDTD.log, replace

use ZLDTDfinal18Jul.dta, clear
***********************************************
```

**********Setting Weights**********

count

set seed 1003002849

sample 46

count

gen pw= 712/328

gen fpc= 712

svyset [pweight=pw], fpc(fpc)

****

**********Demographics**********

svydes

hist AGE , norm

svy: mean AGE

svy: tab EDULEVEL

svy: tab RELIGION

svy: tab MARITAL

svy: tab ETHNICITY

hist LOSYRS, norm

svy: mean LOSYRS

hist LOSMNTS, norm

svy: mean LOSMNTS

****

**********Univariate analysis**********

svy: logistic STI LOSYRS

svy: logistic STI LOSMNTS

svy: logistic STI AFHM

svy: logistic STI LAFHM

svy: logistic STI ALCOHOL

svy: logistic STI AFD

xi:svy: logistic STI i.ALCFREQ

svy: logistic STI ETD

svy: logistic STI DRUGS

svy: logistic STI CANNABIS

svy: logistic STI COCAINE

svy: logistic STI ECSTASY

svy: logistic STI AMPH

svy: logistic STI OPIUM

svy: logistic STI HASH

svy: logistic STI CRYSTAL

svy: logistic STI HEROIN

svy: logistic STI OTHER

svy: logistic STI DRUGSEX

svy: logistic STI SEXCOC

svy: logistic STI SEXECS

svy: logistic STI SEXAMPH

svy: logistic STI SEXOP

svy: logistic STI SEXHASH

svy: logistic STI SEXCRYST

svy: logistic STI SEXHEROIN

svy: logistic STI SEXVIAG

svy: logistic STI SEXOTHER

svy: logistic STI IDU

svy: logistic STI HADSEX

svy: logistic STI AFS

svy: logistic STI CONUSE

svy: logistic STI SEXHAD

svy: logistic STI SEXFEM

svy: logistic STI RFP

svy: logistic STI NRFP

svy: logistic STI CFP

xi:svy: logistic STI i.RPCONUSE

svy: logistic STI CONUSELT

xi:svy: logistic STI i.RFPANAL

xi:svy: logistic STI i.NRPCONUSE

xi:svy: logistic STI i.CPCONUSE

svy: logistic STI CPLT

xi:svy: logistic STI i.CPANAL

svy: logistic STI CIRCUM

svy: logistic STI USECON

svy: logistic STI CARCON

svy: logistic STI AGE

xi:svy: logistic STI i.EDULEVEL

xi:svy: logistic STI i.RELIGION

xi:svy: logistic STI i.MARITAL

xi:svy: logistic STI i.ETHNICITY

****

**********Multivariate analysis*********

svy: logistic STI ETD

svy: logistic STI ETD AFHM

svy: logistic STI ETD AFHM DRUGS

svy: logistic STI ETD AFHM DRUGSEX

svy: logistic STI ETD AFHM CONUSE

svy: logistic STI ETD AFHM CONUSE NRFP

svy: logistic STI ETD AFHM CONUSE CFP

svy: logistic STI ETD AFHM CONUSE CFP ALCOHOL

svy: logistic STI ETD AFHM CONUSE CFP CARCON

svy: logistic STI ETD AFHM CONUSE CFP USECON

svy: logistic STI ETD AFHM CONUSE CFP

****

*******GOF & Diagnostics**********

estat gof

svylogitgof

linktest

collin ETD AFHM CFP CONUSE

coefplot, drop(_cons) xline(1) eform xtitle(Odds ratio)

****

capture log close

APPENDIX A

Approval letters (A1)

- Joint research ethics committee (JREC) approval letter
- Approval letter to use data for secondary data analysis

Questionnaires (A2)

- Questionnaire for Sex workers
- Questionnaire for Long distance truck drivers