BINDURA UNIVERSITY OF SCIENCE EDUCATION
FACULTY OF SCIENCE AND ENGINEERING
DEPARTMENT OF STATISTICS AND MATHEMATICS

Modeling Zimbabwe's Sexual Transmitted Infections Prevalence 2008-2022 (Parirenyatwa Hospital).

By

CHIKODZA BENARDETTE LETPACY

B190721B

A RESEARCH PROJECT SUBMITTED TO THE DEPARTMENT OF MATHEMATICS IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF HONORS BACHELOR OF SCIENCE DEGREE IN STATISTICS AND FINANCIAL MATHEMATICS

Supervisor: DR. T. W. Mapuwei

2023

DECLARATION OF AUTHOURSHIP

I Benardette Letpacy Chikodza declare that this project is my own original work and has not been copied or extracted from previous sources without due acknowledgement of the sources.

12/06/2023

Signature Date

Approval Form

The undersigned certify that they have read and recommend to the Bindura University of Science Education for acceptance of a dissertation entitled “**Modelling Zimbabwe’s Sexual Transmitted Infections Prevalence** ”.

Submitted by **Benardette Letpacy Chikodza**, Registration Number **B190721B** in partial fulfillment of the requirements for the Bachelor of Science Honours degree in Statistics and Financial Mathematics.

Benardette. L Chikodza
12/06/2023

B190721B Signature Date

Certified by

Dr. T. W. Mapuwei 12/06/2023

Supervisor Signature Date

Certified by

Mr. Magodora ……………………. …………….

Chairman of department Signature Date

The Release Form

Registration Number: B190721B

Dissertation Title: Modelling Zimbabwe’s Sexual Transmitted Infections Prevalence.

Year granted: 2022-2023

Authority is given to the Bindura University of Science Education Library and the department of Statistics and Mathematics to produce copies of this Dissertation for academic use only within the University.

Signature of student

Date signed 12/06/2023

DEDICATION

To my parents, a lot of people would have lost their lives if I had gone to a medical school. This is the best doctor I can give you.

ACKNOWLEDGEMENTS

First and foremost, praise and thanks to God, the Almighty for his blessings throughout my learning journey to complete the task successfully.

My gratitude goes to Dr. T. W. Mapuwei for his enthusiasm, insightful comments and hard questions. His advice and guidance helped me all the time of research and writing of this dissertation. Not forgetting all lecturers in department of Statistics and Mathematics for their patience, guidance and generous help.

To my parents, siblings; Fletcher, Becky, Ordette, Andile and Tadisa, for their inspirational messages and words of encouragement and financial support. Also not forgetting my aunty Memory Dambaza, for being my tower of strength through this journey.

To my classmates, for being approachable and for answering my queries of the things I don’t understand about a certain task.

To my friends, for their emotional support and concern especially in cheering me on and helping ma overcome the obstacles encountered in my field study journey.

Last but not least, I want to thank me. I want to thank me for believing in me, I want to thank me for doing all this hard work, I want to thank me for having no days off, I want to thank me for never quitting, I want to thank me for always being a giver and trying to give more than I receive, I want to thank me for trying to do more right and wrong, I want to thank me for just being me at all times.

ABSTRACT

In this study, we assessed whether a person will get sexually transmitted illnesses (STI) for those who underwent testing at Parirenyatwa hospital using factors such as age, marital status, and place of residence.

The main objectives were to identify a suitable model to predict the prevalence of STI’s at Parirenyatwa Hospital, Predicting STI status by residence, HIV status and marital status and to identify factors that influence the prevalence of STI. To identify factors that are important enough to determine whether a person could develop STI or not, we used a logistic regression model. The model's performance was evaluated using various metrics, including accuracy, precision, recall, F1-score, and AUC-ROC score. Visualizations such as, ROC curves and precision-recall curves were performed so as to assess the model's performance. The model showed that those who are married and those who are in committed relationships have low chances of contracting STIs compared to those who are single and separated. The F1 score was used to measure the overall accuracy of the model, while the confusion matrix was used to evaluate the accuracy of the model at predicting true positives, true negatives, false positives, and false negatives.

In conclusion we have seen that those who are HIV negative and marital status (single or not committed in a relationship) significantly increased the likelihood of having a STI in both males and females. Large numbers of those who were affected are from rural areas reason being they are ignorant to STI educative teachings and they have no accesses to condoms ending up having unprotected sex.

Ministry of Health and Child Care should establish more STI screening centers in most rural areas. There should be peer educators at these centers who distribute risk factor knowledge.

Table of content

Contents

List of Tables

List of Figures

CHAPTER 1
INTRODUCTION AND THE BACKGROUND OF THE STUDY

1 Introduction

The most important cornerstones of global pandemic which continues to affect and hinder the normal people's livelihood and cause threats to health is so called the Sexual Transmitted Infections (STI's). Therefore, in Zimbabwe STI's has always been a major challenge which is affecting both the elderly and minor people in our society. This is caused by the ratio of high poverty data timeline, unemployment, lack of knowledge and education in our communities which has triggered to the most outbreaks of STI's. As a result, in emerging nations, STI's has become a conflict and the goal is to come up with measures which does the prevalence's to such diseases from occurring at an alarming rate.

This chapter gives full explanation of the study. It also includes the background of study, statement problem, objectives, research questions, the significance of the study, assumptions of the study, as well as the limitation and the delimitation of the study. Finally, this chapter discusses how the study is carried out and concludes with a summary.

After this background foundation or the introductory chapter there are four additional chapters, making this a five-chapter study. Following the second chapter, a chapter which examines the research literature by reviewing STI's prevalence case at Parirenyatwa in Harare, Zimbabwe. The research methodology chapter, which is the third in the series, focuses on methodologies employed to achieve certain study goals. The fourth chapter on data presentation, analysis and discussion is to lay out a framework for the data process and presentation of research results. The final chapter concentrates on drawing conclusions from the research and giving recommendations based on the results.

1.1 Background of the Study

Globally, each and every year the world continues to note that thousands of the young generation and of course the elderly as well are affected by Sexually Transmitted Infections (STI's). Having observed such trend line of the continuation rise of these diseases, the World Health Organization highlighted that there has been an increase in Human Immune Virus (HIV) since 1980. It is widely known that STI's have a tendency of increasing the infectiousness rate of individuals who are surviving with HIV, by increasing the concentration of viral in their genital tract. Therefore, the Antiretroviral Treatment (ART) lowers the rate of HIV spread through the blood and curb the spreading of infection acquired through STI's. Herpes, chlamydia and vaginal discharge are also infections slowly gaining popularity in the universe and they are known of causing pelvic inflammatory disorder, lowers the fertility rate, introduction of cancerous cells which later own results on claiming human lives. However, there are several traditional ways used to fight the STI's such as completely abstinence, use of condom and educating the marginal society. Amongst the research conducted, showed that at least 2.6 million people in Zimbabwe falls between the age group of 15 to 49 are the vulnerable and exposed to HIV as well as other STI's diseases.

1.2 Statement of the Problem

The prevailing situation of STI's has exposed the country to many factors namely gender, age, marital status, education level and risk of contracting STI's are socio-economic and demographic variables. These several variables play a pivotal role toward the increase of HIV rate and also can be used to find a way to the prevalence of STI's. It is also noted that the upward trend of STI's depicts that the generosity of Zimbabwe are becoming less careful and intolerance hence ending up contracting such diseases.

1.3Objectives

The main objectives of this
study are;

1. To identify a suitable model to predict the prevalence of STI's at Parirenyatwa Hospital;
2. To identify factors that influence the prevalence of sti

1.4 Research Question

In carrying out the research the following questions had to be answered:

1. What model can be incorporated into modelling the STI's prevalence in

Zimbabwe?

2. Will the STI's prevalence fitted model be valid or not?

1.5 Significance of the Study

This study is aimed on addressing problems associated with the STI's prevalence and raise awareness to the people of Zimbabwe. It is also useful for the reader since it improves students' knowledge towards STI's thereby serving as a foundation for those who are keen in learning more about STI's or related topics.

1.6 Assumption of the Study

This study suggests a plausible relationship between STI risk and HIV risk, however intervention studies are still unsatisfactory. This does not rule out a causal relationship, but it does highlight the need for more research into the mechanism of action and the planning and execution of treatments.

1.7 Delimitations

The thesis was delimited to Zimbabwe because this is the place of origin of the researcher. Because there were five dependent variables namely HIV status, place of residence, sex, marital status and age, also one time independent variable namely STI type, therefore the data will be relatively large.

1.8 Limitations of the Study

The researcher faced many challenges in this study. This research study was conducted under the influence of second data sources. Hence, the researcher similarly had to face contains in terms of data accessibility as some of the information was hard to access because of its insensitive nature. As far as the literature review is concerned, the researcher encountered the challenges of choosing appropriate literatures that dwell on Sexually Transmitted Infections (STI's) for countries like Zimbabwe.

1.9 Definition of Terms

Sexually Transmitted Infections

Sexual contact with an infected person is how sexually transmitted diseases (STIs) are spread.

Human Immune Virus

A “deficiency” is when something is weaker than it should be. So, basically, HIV is a virus that attacks the immune system and makes it harder for a person to stay healthy (Royce & Seng, 2021).

Antiretroviral Treatment

Since HIV cannot be cured, nor is there a vaccine to prevent it, the available medications work to minimize the harm HIV does. Overall, the drugs used to treat HIV are referred to as “Anti-Retroviral Therapy” (World Health Organisation,n.d.)

Prevalence

It is described as the quantity of a specific disease or other condition within a specific population at a particular moment.

1.10 Chapter Summary

The chapter outlined the need of forecasting STI’s prevalence to Zimbabwe utilizing a logistic regression indicated from the objectives. As hinted on the statement of the problem, the key aspects which triggered this research is to model STI’s prevalence at Parirenyatwa Hospital. This made the researcher draft the objectives that focus on identifying a logistic regression model, whichever model identified is used to make prediction and forecasting for STI’s as folded by the objectives. As a result, this chapter paved a way for the following chapter which is aimed at reviewing both theoretical literature and related studies.

CHAPTER 2
LITERATURE REVIEW

2 Introduction

Improvements in any field of study come from building on the work of others who came before. According to Webster and Waston (2002), an important undertaking for any academic inquiry is a thorough review of historical literature. This chapter provides relevant material, concept summaries from other investigations, and theoretical literature.

2.1 Related Studies (Empirical Literature)

The majority of sexual interaction is how sexually transmitted infections are communicated. Contrarily, some STIs can also spread by blood, lactation, or childbirth, among other non-sexual methods. As far as STIs go, gonorrhea, genital warts, syphilis, chlamydia, and HIV are the most common. They are particularly harmful for women because, if they are not treated right away, they might result in infertility or cancer. Avoiding unprotected sexual intercourse is the best course of action because they don't exhibit any symptoms.

Numerous research on the transmission of STIs and their varied causes have been carried out throughout Africa. Teenagers have the highest incidence of STIs, according to the Centers for Disease Control and Prevention (2013). This was corroborated by Wilkinson's (2010) study on unrecognized STIs in South African women. Warner and Abdool Karim the researchers employed survey techniques, descriptive statistics, and prediction models in their strategy and analysis. In the Hlabisa region, 13943 women between the ages of 25 and 19 were discovered to have at least one STI. Only 65% of these women sought therapy, even though 52% of them reported symptoms. O. Ohene (2008) conducted a second study in Ghana, focusing on factors associated with STIs in female Ghanaians. They use information from the Ghana Demographic Health Survey, which was carried out in 2003.

A comprehensive model based on the growth of the HIV pandemic was proposed by E. Oster in 2012, and it was replicated in a subsequent study at Harvard University using data patterns, transmission rates, and other epidemic characteristics. The theories that arose emphasized the significance of sexual behavior and transmission predominance in the emergence of epidemics. The study found that when using stimulation models to predict HIV incidence in the United States and Sub-Saharan Africa, the HIV transmission rate is significantly higher for people who have untreated STIs. Sexual behavior was said to be the cause of the higher incidence in Sub-Saharan Africa.

Among young ladies aged 15 to 24 who engaged in sexual activity, Ayo Stephen Adebowade (2013) of Nigeria's University of Ibadon looked into the sociocultural factors for sexually transmitted illnesses. Chi-square and logistic regression were used to evaluate the data. The results showed that STIs had been present for at least a year in females aged 20 to 24. Age, wealth index, and marital status are some examples of socio-demographic characteristics. They came to the conclusion that the wealth index and knowledge of HIV and AIDS were significant predictors of STI acquisition. The risk of STI transmission can be reduced by distributing condoms, promoting the benefits of abstinence, and raising awareness about HIV and AIDS.

Important Definitions

The following are the most important definitions which we will see in this writing:

1. Logistic Regression: This statistical analysis model employs a logistic function to simulate the behavior of a binary response variable.
2. Odds Ratio: These are ratios of proportions for two possible outcomes
3. Sexual Transmitted Infections: These are infections that are transferred from one person to another sexually.
4. Binary Data: It is data that is represented by ones and zeros. It can also be categorized into two groups.

Theoretical Review

The theoretical framework is the framework upon which the hypothesis of a research study can be placed or supported. The theoretical framework introduces and describes the underlying theory of the research topic being investigated. Themes, their definitions, and references to pertinent academic works that were used to develop a specific study make up this section. The theoretical framework must show an understanding of relevant theories and ideas that are related to both the current topic and the study's more broad knowledge areas. The theoretical structure of this study is based on the following ideas:

i. Regression analysis
ii. General linear models
iii. Linear models
iv. Generalized linear models
v. Logistic regression model
vi. Binary logistic regression
vii. Model selection

2.2 Regression Analysis

The approach of regression analysis is employed to investigate the relationship between two or more variables. It focuses on the development of models in which the independent or explanatory characteristic is used to describe a single variable known as the dependent or response variable (represented by Y). One or more response features may be reliant upon the dependent feature.

2.3 General Linear Model

This statistical model has formula $Y = X\beta + \epsilon$

Assumptions of general linear models

1. X is a non-random vector
2. Y and ϵ are independent random vectors
3. Error terms ϵ have constant variables
4. Error terms are normally distributed and have zero mean at all X values.

2.4 Linear Models

Linear regression model for *n* observations $_1$, $_2$... $_n$,is given in the form

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_2 + \cdots + \beta_k X_{ki} + e_i \qquad (2.4.1)$$

In this model , the dependent feature for the ith observant , $_i$ = 1,...,n , is linearly dependent on the values of the k independent variables , $x_i,\ldots,x_k$, through the parameters $\beta_0,\ldots,\beta_k$. It is assumed that error terms , which indicate residual variation , have mean of zero and a constant variance. Likewise , it is anticipated that the variance of independent variables will remain constant. $E[Y_i] = \beta_0 + \beta_1 X_{1i}$ gives the expectation of Y , and $Var[Y] = \sigma^2$. A group of values that are factors are taken into account by linear models , which also include qualitative variables known as factors. General linear models is not compatible for binary data, so we as a result , we applied general linear models.

2.5 Generalized Linear Models

In GLIM the connect function is employed. The link functions in generalized linear models connect the mean response to the model's linear indicators. Assuming that p is the ratio of subject's reaction to a stimulus intensity, g(p) = Xβ is a generalized linear in terms of X.

2.5.1 Regression with Binary Response

Binary has two possible outcomes which can be presented by 1 and 0, therefore let Y_i be the binary. The aim is to obtain the category set of forecasting variables that is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{ip} + e_i \qquad (2.5.2)$$

E[Y_i] = p_i for i = 1,...,n and Y_i ranges between 0 and 1 ,which is the proportion of observations at all dependent variables. $P_i = P(Y_i = 1)$, implying that $1 - P_i = P(Y_i = 0)$, for i = 1, 2, ... , n , which means that n distinct data points , there is n probability of which is a bernoulli distribution parameter.

2.5.2 Bernoulli Distribution

It is the discrete probability of a random variable with success and failure values 0 and 1 respectively. It is an example of binomial distribution. Assume Y is a random variable with one of the two outcomes: success or failure. If P(Y=Success)=p and P(Y=Failure)= 1-p, then Y follows the Bernoulli distribution with parameter p given by :

$$p(x) = p^x(1-p)^{1-x} \qquad (2.5.2)$$

2.5.3 Logistic Transformation

Commonly used in medical or clinical field because we will be modelling odds. Instead of modelling p we model log (p/1-p) which we write as logit(p).

$$\log p/(1-p) = \text{logit}(p) = \beta_0 + \beta_0 X_{1i} + \cdots + \beta_p X_{pi} \qquad (2.5.3)$$

2.6 Logistic Regression Model

For linear discriminant analysis, Joseph Berkison created the logistic regression model in 1944 as a replacement for Fisher's (1936) classification approach. The best regression method to utilize when the dependent variable is dichotomous (binary) is logistic regression. The logistic regression is a predictive analysis, just like other regression studies. One way to define and explain the relationship between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables is to utilize logistic regression. It is used in social science and medical or clinical sciences. When dealing with three or more outcomes, it can be multinomial; when dealing with two outcomes, it can be binomial. Binary logistic regression has two possible codes: 0 and 1, where 0 denotes failure and 1 denotes success.

In general, the results of the independent variables are used to determine whether a logistic regression would be successful. Additionally, it uses one or more independent variables that could be categorical or continuous. The Bernoulli distribution is also utilized, which can be used to predict binary events rather than continuous ones. In logistic regression, the dependent variable, which is the natural logarithm of the probabilities, can be used to generate continuous criteria as the modified form of the dependent variable. We contend that the logit transformation is the link function on which logistic regression is carried out, despite the fact that the response variable in logistic regression is binomial and the logit is the continuous factor on which it is based.

2.6.1 The Purpose of Logistic Regression

Logistic regression is used for categorical data collecting. When our data is dichotomous, we use logistic regression rather than linear regression. It is possible to automatically create dummy variables with logistic regression, which makes it simpler to use. Odd ratios are used to display the findings of logistic regression, which is primarily used to forecast categorical interactions. Its second significance is that it demonstrates how closely associated and interconnected the variables are to one another.

Assumptions of logistics regression differs from linear regression in that it does not make some fundamental assumptions that linear and general linear models had hold so close. These are the assumptions:

1. Regression and the response variable should not have linear relationship

2. Errors should not follow normality.
3. Variance of error term is not constant, there is no homoscedasticity.
4. In logistic regression measuring the response variable on an interval or ratio scale is not applicable.

However logistic regression shares some assumptions with linear regression such as:

1. Linearity of regresser variables and log odds – logistic regression is the predicted on the assumption that the independent variables are linear and the probabilities are logarithmic. The independent variables must be linearly connected to the log odds, since this approach does not require a linear relationship between the dependent and independent variables
2. Absence of multicollinearity – to apply logistic regression, the independent variables must have low level of multicollinearity. This shows that the independent variables should not be too related.
3. Observation independent – observations should be independent of one another when using logistic regression. In other words, observations should not be based on repeated measurement.
4. Assumptions of large sample size – lastly, logistic regression often necessitates a large sample size. A common rule of thumb is that each independent variable in your model should include at least 10 occurrences with the least likely outcome.

The logistic regression model is given by

$$\ln\left[\frac{\hat{p}}{1-\hat{p}}\right] = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

$$\hat{p} = \frac{\exp\ (\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}{1\text{-}\exp\ (\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)} \qquad (2.6.5)$$

$$\text{logit}(p_i) = (p_i / 1\text{-}p_i)$$

2.6.2 Limitations of Logistic regression model

Logistic regression needs random independent sampling and linearity between X and the logit, so it does not need multivariate normal distribution. At the extremes of the distributions, the model is less likely to be accurate and more likely to be correct towards the center. Although P(Y=1) maybe predicted for all possible value combinations, not every possibilities may occur in the population. Models run risk of becoming a mislead when the crucial variables are left out. On the other side, including unimportant elements may lessen the effects of more important ones. Although multicollinearity may not lead to skewed results, it does increase standard errors for coefficients and make it more challenging to statistically distinguish the distinct contribution of overlapping variables.

More data is always important. Unstable models are often as a result of small sample size. Focus out for outliers who might cause correlations to become distorted. In small samples, some value combinations may be relatively weakly represented in correlated variables. When estimates are based on cells with modest expected values, they are unstable and lack power. Perhaps small categories can be meaningful collapsed. Plot the data so to ensure if the model is correct. Is it necessary to interact? Take caution not to misinterpret odds ratio for risk ratios.

2.7 Binary Logistic Regression

2.7.1 Statistical Modeling

The primary goal of the modelling is to provide a mathematical representation of the interaction between an observed variables and a set of independent variables. These models are employed for a variety of purposes, including

a. Forecasting the dependent variables based on the predictor variables' values.
b. Comprehending how the predictor variables affect or connect to the dependent variable.

2.7.2 Odds Ratios

Definition Odds: is defined as the probability of an event happening divided by the probability that it won't happen.

Retrospective study

It is an etiological study that compares with (cases) and without (controls) a certain disease.

Confounding Variables

Collett (1991) define it as a variable that completely or partially accounts for the apparent connection between a disease and an exposure factor. The odds ratios defined as the frequency with which certain occurrences occur. Odds ratios are used in statistics to assess how the existence of A is connected to the presence or absence of B in a particular population. It is also used to estimate the likelihood of an interest result occurring when a variable of an interest is exposed. The first result occurs a number of times for every single occurrence of the second result, this is commonly stated as a:b. Odds are calculated by dividing the probability by one minus probability.

One of the three primary methods in epidemiology for determining the relationship between outcome and exposure is confounding variable. The other two basic approaches of assessing association are the Risk Relative (RR) and Absolute Risk Reduction (ARR). A disease’s relative risk is a measure of how likely individual exposed to a certain cause is to develop a disease compared to someone who is not exposed. Relative risk is similar to odds, except instead of odds, utilizes probabilities and its mainly used in medical studies.

But since this study is retrospective, it captures the use of odds ratio. For each unit increase in the predictor, the odds ratio shows how the likelihood of being a member of the target group changes. It is calculated using the regression coefficient of the exponent. In the Statistical Package for Social Sciences, the probabilities are computed and displayed as exp (SPSS). It shows how much altering the related the odds ratio is influenced by the measure by one unit. It enhances the likelihood of a successful outcome if it is greater than one. Each increase in the predictor lowers the odds of the result occurring if it is less than one. If the odds are 1, there is no connection between the variable and the odds.

2.8 Model Estimation

You can estimate the model's unknown parameters to fit a model to a set of data that you should already have. The two most common strategies are maximum likelihood and least squares. Point estimation and interval estimation are the two techniques available for estimating unknown parameters. The methodologies listed below are used for point estimate.

a) Maximum Likelihood estimation
b) Methods of moments
c) Methods of least squares
d) Judgmental methods

2.8.1 Maximum Likelihood Function

Model parameter estimations are provided. Given a defined collection of data and fundamental model, the maximum likelihood model chooses the set of model parameters and maximizes the likelihood function. Consider a sample of independent observations with the same distribution x_i,.., x_n from a distribution with an unknown probability density function $f_0(.)$. However, it is distilled into the fact that the function f_0 is a member of a particular family of distributions (where is a vector of parameters from this family), known as the parametric model, and that $f_0 = f(./\theta_0)$ the value of θ is unknown and is referred to as the true values of parameter vectors.

2.8.2 Methods of Moments

Assume y_i, … ,y_n are independent observations with $E(Y_i) = \sum\beta_j X_{ji}$ and $Var(y_i) = \ ^2$ for all i = 1,2, … , n. The least square estimates of the unknown parameters in the model are values of $\beta_0 \ldots \beta_n$, which minimizes the error of squared deviations of the observations from their predicted value given by

$$S = \sum\{Y_i - E(y_i)\}^2 = \sum(y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_r X_{ri})^2 \qquad (2.8.6)$$

By differentiating S with respect to each of the unknown parameters and setting derivatives to zero, the least squares estimate is achieved.

2.8.3 Fitting the Logistics Regression Model

To estimate the parameters β_0 and β_1 after obtaining the logistic regression model, we must fit the model to a set of data. We stated that the straight line fitting the data in linear regression might be produced by reducing the distance between each dot on a plot and the regression line. In order to avoid negative differences, we really minimize the sum of the squares of the distance between the dots and the regression line. The least sum of squares approach is what is used in this situation. By using the least sum of squares, we determine the values of β_0 and β_1.

The logistic regression process is challenging. The maximum likelihood technique is what it is called. Maximum likelihood will offer 0 and 1 values that increase the likelihood of finding the data set. Iterative calculation is necessary, and most computer software makes it simple to measure. The likelihood function is used to calculate the likelihood of getting the data given the unknown parameters (β_0 and β_1).

The probability varies from 0 to 1.

The logarithm of the likelihood function is useful in practice. The log-likelihood function will be utilized for inference testing when contrasting several models. The natural log of any number less than 1 is negative, hence the log probability is between 0 and −∞ (it is negative). As follows is the calculation of the log-likelihood:

$$l(\theta;x_i, \ldots, x_n) = \sum_{i=1}^{n} \ln \quad (f(x_i\theta)) \qquad (2.8.7)$$

The likelihood varies on the unknown success probabilities P_i which depends on through the equation $p_i = \exp\left(\beta_0+\beta_1X_{1i}+\cdots+\beta_pX_{pi}\right)/1 + \exp \quad (\beta_0 + \beta_1X_{1i} + \cdots + \beta_pX_{pi})$

Logarithm likelihood function is given by

$$L(\beta) = \prod_{i=1}\binom{x}{y}p_i^{y_i}\left(1\text{-}p_i\right)^{n_i\text{-}y_i} \qquad (2.8.8)$$

Derivatives with respect to (r+1) unknown parameters are given by

$$\frac{\partial \log L(\hat{\beta})}{\partial\hat{\beta}} = \sum y_i x_{ji} - \sum n_i x_{ji} e^{ni}\left(1+e^{ni}\right)^{-1}$$

It is possible to numerically solve a set of r + 1 nonlinear equations with the unknown parameters $\hat{\beta}_j$ by resolving the derivatives of $\hat{\beta}$ and equating them to zero. Two ways for evaluating these nonlinear equations which are

a. Newton Raphson
b. Fisher's methods of scoring

2.8.4 Newton Raphson

Newton Raphson is a generalized model fitting algorithm. The log likelihood function r + 1 derivatives with respect to $\beta_0, \ldots, \beta_p$ are also known as efficient scores and are indicated by vector

$$\begin{pmatrix} \dfrac{\partial \log L(\hat{\beta})}{\partial \hat{\beta}_0} \\ \dfrac{\partial \log L(\hat{\beta})}{\partial \hat{\beta}_1} \\ \vdots \\ \dfrac{\partial \log L \quad (\hat{\beta})}{\partial \hat{\beta}_r} \end{pmatrix} \qquad (2.8.9)$$

Let H (β) be a matrix of second partial derivatives of logL, with the (j, k)th element of H(β) equal to dlogL (β)/$\beta_J\beta_{K,}$ where j = 1, 2, 3, … , r and k = 1,2, …, r. H (β) is the Hessian matrix. Taking into account $U(\hat{\beta})$ assessed at the maximum likelihood estimate of β, β^ uses Taylor series to extend $U(\hat{\beta})$ and β* where β* is close to β^ and we have:

$$U\left(\hat{\beta}\right) = U(\beta^*) + H(\beta^*)(\hat{\beta} - \beta^*) \qquad (2.8.9)$$

2.8.5 Fisher Scoring Method

This method uses the information matrix rather than the Hessian matrix, the (j, k)th element of which is $-E\text{-}E\left[\dfrac{\sigma^2 \log L(\beta)}{\sigma\beta_j\beta_k}\right]$ for j = 0, 1, …, r and k = 0,1…r and this matrix is indicated by I(β) by using the iterative methods outlined above to replace the Hessian matrix. In case of the logit techniques, they all approaches to the identical maximum likelihood estimate of β due to different iterative techniques which provide different results in general.

2.9 Chapter Summary

This chapter explains other authors' views regarding the methods used to model STI's prevalence. The following chapter focuses on the research methodology where the techniques used to collect analyze data are explained.

CHAPTER 3
RESEARCH METHODOLOGY

3 Introduction

The methods for applying linear logistic regression to the model of sexually transmitted diseases based on age, marital status, HIV status, residential location, and gender will be covered in this chapter. The descriptive statistics that will be applied must also be considered.

3.1 Research design

The primary method that adequately addresses the queries raised in the first chapter is research design. A strategy, layout, and method of examination that is thought of as a manner of acquiring answers to research questions, is also a research design. Research design is a format and structure utilized when conducting research, according to Best (1993). In light of these viewpoints, a research design is a diagram of the procedures and techniques used to gather and analyze data. The research project is held together by its research design, which acts as glue. The patient-level data were prospectively gathered throughout the trial period, despite being examined retrospectively. This study employed a quantitative research approach to quantify the issue by producing numerical data or data that could be converted into useful statistics. Quantitative research makes use of quantifiable data to establish facts and identify trends in study. This study methodology produced unbiased, statistical, and logical results.

3.2 Data source

The data used in the study is a secondary data sourced from Parirenyatwa Hospital, which is a tertiary level hospital located in Harare, Zimbabwe. The data used in the study was randomized and anonymized, which means that any personal information that could be used to identify individual patients was removed to protect their privacy. The data covers a period of 2018 to 2022, which provides a sufficient time frame to assess the prevalence of STIs in Zimbabwe. By using data from a reputable hospital, the study is able to draw conclusions that are representative of the population in Zimbabwe and can be used to inform STI prevention and control strategies in the country.

3.3 Data analysis

The study uses data on age, sex, STI type, HIV status, marital status, and residence to build the model. The logistic regression model will be evaluated using quality and precision checks, including the F1 score and confusion matrix. The F1 score will be used to measure the overall accuracy of the model, while the confusion matrix will be used to evaluate the accuracy of the model at predicting true positives, true negatives, false positives, and false negatives. The distribution of STIs among males and females analyzed using descriptive statistics, including histograms and correlation matrices. The researcher also calculated mean, standard deviation, minimum and maximum for continuous variables

3.4 Mode of Analysis

Python package is going to be used for data analysis. Python is a popular programming language that is widely used in data analysis and machine learning. In the context of the study, Python was used to conduct the logistic regression analysis to model the prevalence of STIs in Zimbabwe. One of the key benefits of using Python for data analysis is its versatility. Python has a wide range of packages, such as numpy, pandas, and scikit-learn, that are specifically designed for data analysis and machine learning. These packages provide a range of tools for data manipulation, visualization, and modeling, which make the data analysis process faster and more efficient.

Another benefit of using Python is its ease of use. Python has a simple and intuitive syntax that is easy to learn, even for individuals with limited programming experience. Python's readability and simplicity also make it easier to collaborate with other researchers and share code.

Finally, Python's popularity and widespread use in the data analysis and machine learning communities mean that there are many resources available for learning and troubleshooting. This includes online tutorials, forums, and documentation, as well as a large community of developers who can provide support and guidance. Overall, using Python for data analysis offers a range of benefits, including versatility, ease of use, and a supportive community, which make it an excellent choice for conducting data analysis in the research context.

3.5 Model selection

Model selection is the process of selecting a statistical model from a collection of models. It is important to compare the performance of different models when there are many models available using different statistical markers. The model's performance will be evaluated using various metrics, including accuracy, precision, recall, F1-score, and AUC-ROC score. Visualizations such as, ROC curves and precision-recall curves will be performed so as to assess the model's performance.

3.6 Chapter Summary

The chapter has explained the methodology that is going to be used. Excel will be used to collect data and another data is going to be analyzed in SPSS. The following chapter will focus on data analysis and interpretation of results.

CHAPTER 4
DATA ANALYSIS AND PRESENTATION OF FINDINGS

4 Introduction

This chapter will focus on the use of logistic regression to model the prevalence of STIs in Zimbabwe. The chapter will begin by providing an overview of the data used in the study, including the source and sample size.

4.1.0 Data profiling

Data profiling is an important step in any data analysis process as it provides a comprehensive understanding of the data and its structure. Data profiling involves examining the data to identify patterns, anomalies, and inconsistencies. It helps to ensure that the data is of sufficient quality and completeness to support the analysis and that any issues are identified and addressed early in the process.

In the context of this study, data profiling was conducted in Python to gain a better understanding of the STI prevalence data from Parirenyatwa Hospital. The data profiling process involved examining the data through descriptive statistics, such as histograms and correlation matrices, to identify any patterns or anomalies. This helped to identify any issues with the data, such as missing values or outliers, which were then addressed before conducting the logistic regression analysis.

Data profiling involves several subtopics, including:

a) Data types: This involves identifying the data types of the variables in the dataset, such as numerical or categorical data.
b) Data completeness: This involves checking percentage and value of missing data.
c) Data quality: This involves identifying any data errors, such as inconsistencies or outliers, that may affect the analysis.
d) Data distribution: This involves examining the distribution of the data to identify

any patterns or anomalies.

e) Data relationships: This involves examining the relationships between variables in the dataset to identify any correlations or dependencies.

Overall, data profiling is important in ensuring the quality and completeness of the data used in any analysis. By conducting a thorough data profiling process, researchers can ensure that the analysis is based on high-quality and reliable data, which can lead to more accurate and robust conclusions.

In this study, data profiling was conducted in Python using various functions and methods provided by the numpy and pandas packages. For instance, the describe() method was used to generate summary statistics for the dataset, including measures of central tendency and dispersion. Histograms and scatter plots were also created to visualize the distribution of the data and identify any patterns or outliers. Additionally, correlation matrices were used to examine the relationships between variables in the dataset. By conducting these data profiling steps, the researchers were able to gain a deeper understanding of the data and ensure the quality and completeness of the dataset used in the logistic regression analysis.

4.1.0 Summary statistics

Table 4. 1 Summary Statistics of variables

Column1	Column2	Column3	Column4	Column5	Column6
	age	sex	. . .	marital stat	residence
count	2800	2800	. . .	2800	2800
mean	32.107143	1.432857	. . .	3.520714	1.441786
std	9.579584	0.49556	. . .	1.596352	0.496688
min	15	1	. . .	1	1
25%	25	1	. . .	2	1

50%	30	1	. . .	4	1
75%	37	2	. . .	5	2
max	66	2	. . .	5	2

These summary statistics present information about a dataset which contains 2,800 observations on seven variables. Here is what we can infer from the statistics:

Age: The average age of the individuals in the dataset is 32.1 years, with a standard deviation of 9.6 years. The youngest person in the dataset is 15 years old, and the oldest is 66 years old.

Sex: The variable 'sex' is coded as 1 for male and 2 for female. The majority of the individuals in the dataset are male, as the mean is 1.43.

Marital Status: The variable 'marital_status' is coded as 1 for married, 2 for divorced, 3 for separated, 4 for widowed, and 5 for single. The majority of the individuals in the dataset are either married or single, as the mean is 3.52.

Residence: The variable 'residence' is coded as 1 for urban and 2 for rural. The majority of the individuals in the dataset live in urban areas, as the mean is 1.44.

4.1.1 Variables in the dataset

The dataset was viewed the first 5 rows so as to inspect the variables in it and the number of rows and columns. The output was as follows; shown as table 4.1.1

Table 4.1. 1 variables in data set

Column1	Column2	Column3	Column4	Column5	Column6	Column7
code in py	age	sex	sti type	HIV status	marital stat	Res

0	33	1	1	1	3	1
1	46	1	0	2	4	1
2	36	2	2	1	5	2
3	37	1	1	2	5	1
4	27	2	2	2	5	1
Number of rows = 2800 Columns= 7						

From the output on table 4.1.1 we can note that the dataset has 2800 rows and 7 columns.

The codes in the have been explained on part 4.1

4.1.2 Data types

Table 4.1. 2 data types

Column1	Column2
data type	Int
age	int64
sex	int64
sti type	int64
sti status	int64
hiv status	int64
marital stat	int64
residence	int64

The dataset contains seven variables, including age, sex, sti_type, sti_status, hiv_status, marital_status, and residence. Each variable has a data type of int64, which indicates that they are integer values with a maximum size of 64 bits.

The dtype: object line refers to the data type of the Pandas Series object that contains the variables. The dtype attribute of the DataFrame refers to the data type of the object, which is object in this case.

Overall, this output provides a summary of the data types in the dataset and is useful for ensuring that the data is in the correct format for analysis. By identifying the data types of the variables, researchers can ensure that the data is appropriate for the analysis and that any necessary data transformations are applied. For example, if a variable was incorrectly classified as an integer when it should be a categorical variable, it would need to be converted to the correct data type before conducting further analysis.

4.1.3 Missing Data

It is important to check for missing data in a study modeling STI prevalence in Zimbabwe, and to handle missing data appropriately to ensure that the results are accurate, reliable, and unbiased.

Data was checked its missingness in python and the output were as follows;

Table 4.1. 3 missing data

data type	missing value
Age	0
Sex	0
sti type	0
sti status	0
marital stat	0
Residence	0

The results have shown that there are no missing values in the dataset for any of the variables: age, sex, sti_type, sti_status, hiv_status, marital_status, and residence. This means that the dataset is complete and there is no need to handle missing values before proceeding with the analysis.

However, it is important to keep in mind that there may still be other data quality issues that need to be addressed, such as outliers, inconsistencies, or errors in the data. It is recommended to perform exploratory data analysis and data cleaning to identify and address any such issues before proceeding with the analysis.

4.1.4 Visualising the variables

Visualizing variables through graphs is an important step in data analysis that can help to identify patterns, communicate findings, evaluate assumptions, and guide further analysis. The plot below was used to visualize the distribution of variables in the dataset.4.

Histograms of visualizing variables

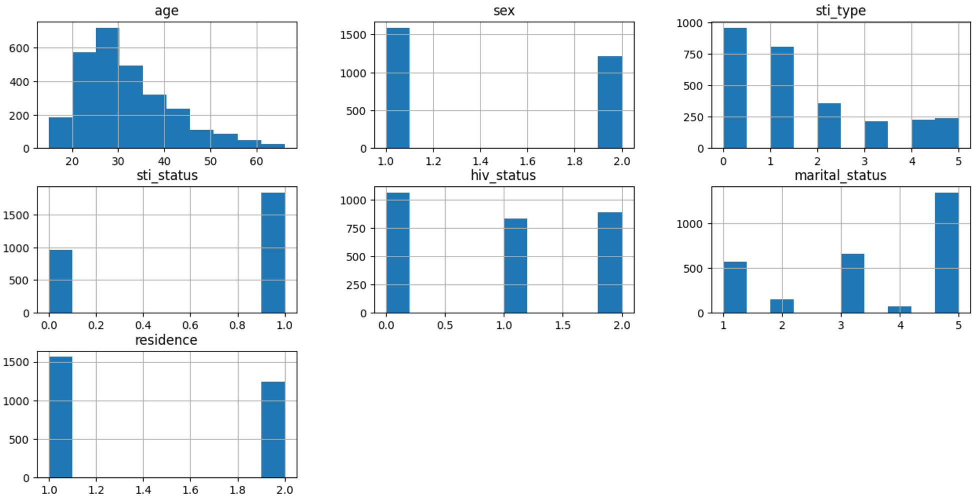


Fig 4.1.4 1

4.1.5Visualizing age

Visualizing the age distribution through a histogram with a normal curve fitted can provide important insights into the shape of the distribution, as well as the skewness and kurtosis of the age variable. Here are some explanations of the importance of this visualization, the outcome from python and how to interpret it:

Graph of visualizing age

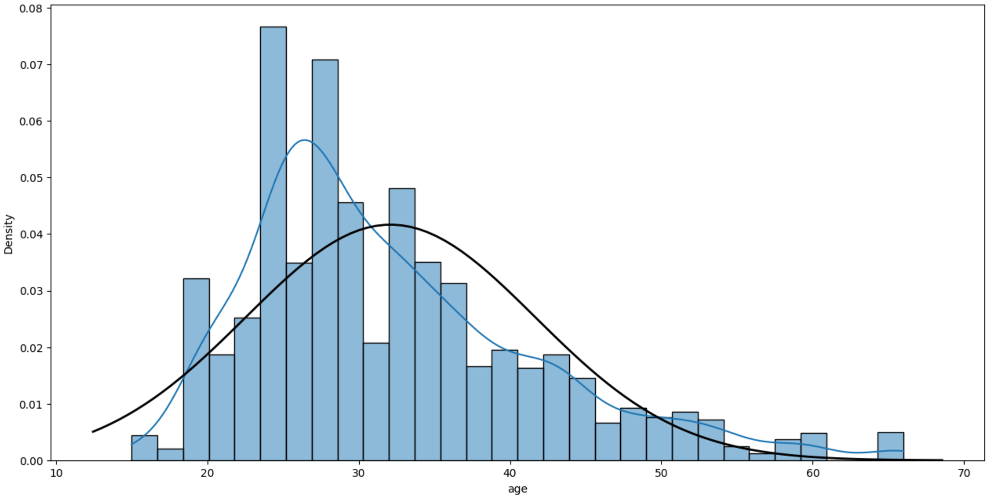


Fig 4.1.5 1

Identifying the shape of the distribution: A histogram with a normal curve fitted can help to identify the shape of the age distribution, which can provide insights into the central tendency and variability of the variable. For example, if the distribution is roughly symmetrical, it suggests that the mean and median of the age variable are similar, and that the data points are evenly distributed around the central tendency.

Checking for normality: A histogram with a normal curve fitted can also help to check if the age variable follows a normal distribution. If the distribution is approximately bell-shaped and the normal curve fits well, it suggests that the age variable is normally distributed, which is important for some statistical analyses.

Interpreting skewness: Skewness is a measure of the asymmetry of a distribution. If the histogram is skewed to the right, it suggests that the distribution has a long tail on the right side, and that the mean age is higher than the median age. Conversely, if the histogram is skewed to the left, it suggests that the distribution has a long tail on the left side, and that the mean age is lower than the median age.

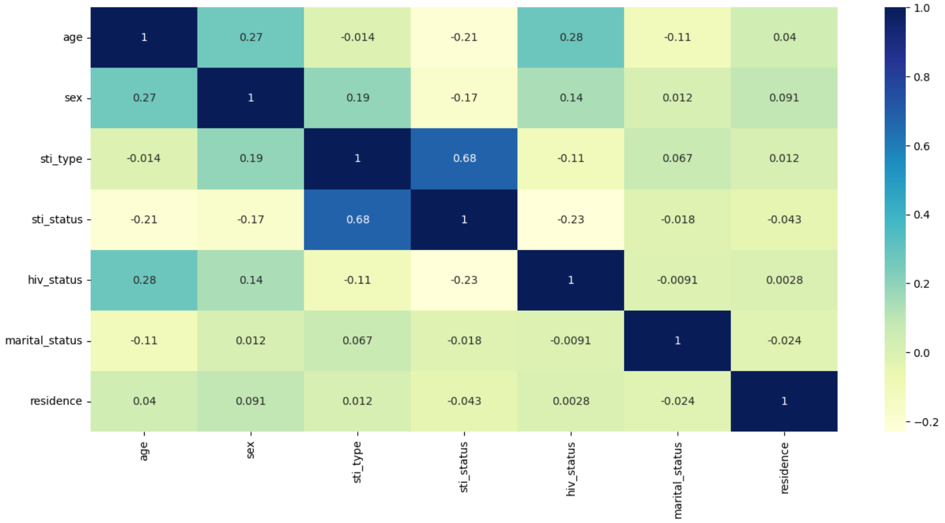
Interpreting kurtosis: Kurtosisis a measure of the peakedness of a distribution. If the histogram has high kurtosis, it suggests that the distribution has a sharp peak and heavy tails, which means that the age variable has more extreme values than a normal distribution. Conversely, if the histogram has low kurtosis, it suggests that the distribution is flatter and more spread out than a normal distribution.

In summary, visualizing the age distribution through a histogram with a normal curve fitted is important for identifying the shape of the distribution, checking for normality, and interpreting skewness and kurtosis. These insights can inform further statistical analyses and help to understand the distribution of the age variable in the dataset.

4.1.6 Visualizing the correlation of variables

Visualizing the correlation of variables is an important step in data analysis that can help to identify relationships, evaluate assumptions, guide feature selection, and communicate findings.

The correlation plot of variables



	age	sex	sti_type	sti_status	hiv_status	marital_status	residence
age	1	0.27	-0.014	-0.21	0.28	-0.11	0.04
sex	0.27	1	0.19	-0.17	0.14	0.012	0.091
sti_type	-0.014	0.19	1	0.68	-0.11	0.067	0.012
sti_status	-0.21	-0.17	0.68	1	-0.23	-0.018	-0.043
hiv_status	0.28	0.14	-0.11	-0.23	1	-0.0091	0.0028
marital_status	-0.11	0.012	0.067	-0.018	-0.0091	1	-0.024
residence	0.04	0.091	0.012	-0.043	0.0028	-0.024	1

Fig 4.1.6 1

4.2 Assumptions of logistic regression

In order to model the link between a binary outcome variable (such as the presence or absence of STI) and one or more predictor variables, the statistical method of logistic regression is often used. Logistic regression, like every statistical method, is valid only under particular conditions. The main tenets of logistic regression are listed below.

Binary outcome variable: Logistic regression assumes that the outcome variable is binary, meaning that it takes only two possible values (e.g., 0 or 1, presence or absence of STI). If the outcome variable is not binary, logistic regression may not be an appropriate technique. In our dataset the variable is sti_status which indicates the presents of sti or its absence.

Independence of observations: The observations are thought to be independent of one another according to the logic regression. It follows that the presence or absence of STI in one person has no bearing on the presence or absence of STI in another.

The correlation matrix

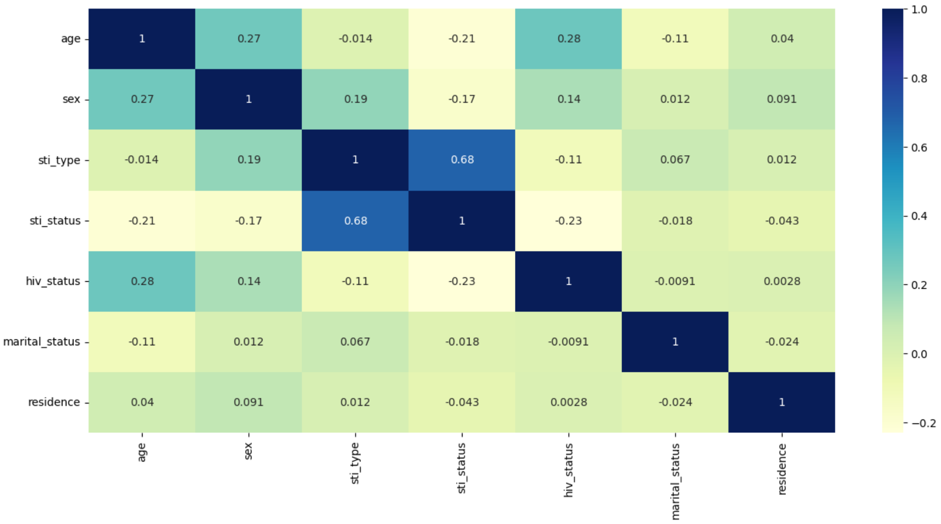


Fig 4.2 1

High correlations indicate that the predictor variables are not independent of each other which violates the assumption of independence of observations. So in this case sti status and sti type are highly correlated so sti type will be removed for further analysis that fitting the logistics regression.

No multicollinearity: Logistic regression presupposes that there is no multicollinearity among the predictor variables. This indicates that the predictor variables are not highly correlated with each other. The vif in conjunction with correlation matrix was used to check this and the output was as follows;

Table 4. 2 VIF

Column1	Column2	Column3
	VIF Factor	Features
0	41.631077	const
1	1.197518	age
2	2.404731	sti stat
3	1.313865	sex
4	2.330617	sti type
5	1.123839	hiv stat
6	1.034416	marital stat
7	1.010821	residence

The table 4.2 shows the Variance Inflation Factor (VIF) for each feature in a logistic regression model. The VIF measures the degree of multicollinearity between the independent variables in a model. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other, which can lead to inaccurate estimates of the relationship between each independent variable and the dependent variable.

In this case, the const feature has a very high VIF of 41.63, which suggests that it may be a problematic variable in the model. The other independent variables in the model have VIF values ranging from 1.01 to 2.40. Generally, a VIF value of 1 indicates that there is no multicollinearity between the independent variables, while a VIF value greater than 5 or 10 suggests a high degree of multicollinearity. Therefore, the values in this table 4.2 suggest that there is relatively low multicollinearity between independent variables in the model.

Overall, the low VIF values for the independent variables suggest that multicollinearity is not a major concern in this model. If there is multicollinearity, it can be difficult to determine the unique contribution of each predictor variable to the outcome variable, and the estimates of the regression coefficients may be unstable.

Adequacy of sample size: Logistic regression assumes that the sample size is adequate for the number of predictor variables included in the model. A general rule of thumb is that there should be at least 10 observations for each predictor variable. From the data profiling we can note that each variable has 2800 observations which is sufficient for modeling sti prevalence. If the sample size is too small, the estimates of the regression coefficients may be unstable, and the model may not generalize well to new data.

Absence of influential outliers: Logistic regression assumes that there are no influential outliers in the data. Influential outliers are data points that have a large effect on the estimated regression coefficients. The following plot for age was used to check for outliers.

Checking for Outliers

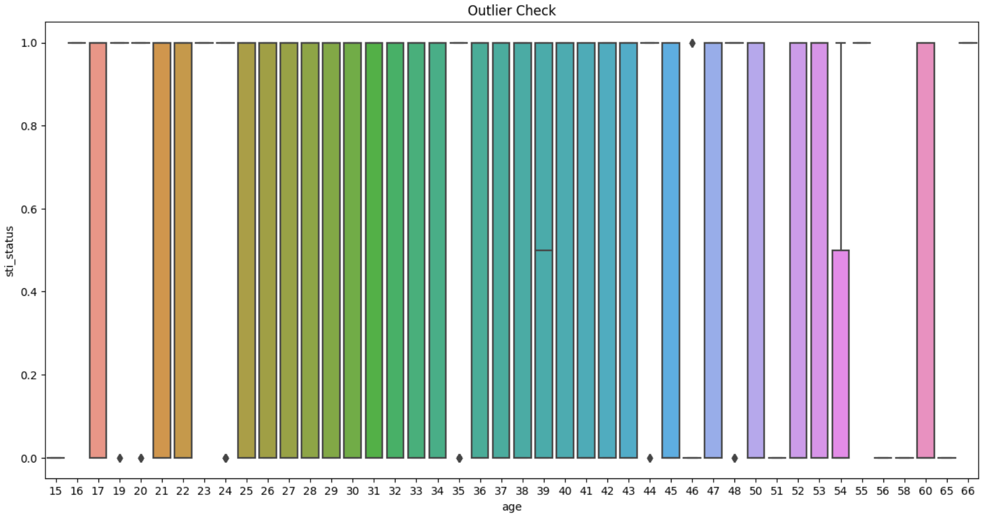


Fig 4.2 2

Therefore, when modeling the prevalence of STI in Zimbabwe using logistic regression, it is necessary to check the assumptions of the model to ensure that they are valid. Violation of any of these assumptions can lead to biased or unstable results, and may require the use of alternative modeling techniques or data transformations.

4.3Model fitting

Logistic regression is a statistical technique used to model the relationship between a binary outcome variable and one or more predictor variables. In the case of modeling STI prevalence in Zimbabwe, the outcome variable was sti_status (e.g., STI positive or negative), and the predictor variables were age, sex, sti_type, hiv_status, marital_status and residence. 4.3.1The code was used for fitting the model.

```
# Split data into features and sti_status
X = data.drop("sti_status", axis=1)
y = data["sti_status"]

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Fit logistic regression model
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
# Print model coefficients
print("Intercept:", logreg.intercept_)
print("Coefficients:", logreg.coef_)
```

The code provided is how the logistic regression model was fitted to the data using Python and the scikit-learn library.

The intercept represents the log-odds of the outcome variable when all predictor variables are equal to zero, while the coefficients represent the change in log-odds of

the outcome variable associated with a one-unit increase in each predictor variable. These coefficients can be used to make predictions on new data.
yThe outputs from this were;

Table 4. 3 Parameter Estimates

Column1	B
Intercept	3.23312143
Age	-0.0293972
Sex	-0.55899342
hiv status	-0.47940932
marital stats	-0.05638116
Res	-0.09324959

Here's what each of these coefficients mean in the context of the variables:
Intercept: The intercept is the log-odds of having an STI when all of the predictor variables are equal to zero. In this case, the intercept is 3.23312143, which means that the log-odds of having an STI for a person with all predictor variables equal to zero is 3.23312143.

Age: The coefficient for age is -0.0293972, which means that for every one-unit increase in age, the log-odds of having an STI decrease by -0.0293972, holding all other variables constant.

Sex: The coefficient for sex is -0.55899342, which means that males are more likely to have an STI than females, as indicated by the negative sign. Specifically, the log-odds of having an STI for males is -0.55899342 higher than for females, holding all other variables constant.

HIV status: The coefficient for HIV status is -0.47940932, which means that HIV-positive individuals are less likely to have an STI than HIV-negative individuals. Specifically, the log-odds of having an STI for HIV-positive individuals is -0.47940932 lower than for HIV-negative individuals, holding all other variables constant.

Marital status: The coefficient for marital status is -0.05638116, which means that being married or in a committed relationship is associated with a lower likelihood of having an STI than being single or not in a committed relationship. Specifically, the log-odds of having an STI for individuals who are married or in a committed relationship is -0.05638116 lower than for individuals who are single or not in a committed relationship, holding all other variables constant.

Residence: The coefficient for residence is -0.09324959, which means that living in a rural area is associated with a lower likelihood of having an STI than living in an urban area. Specifically, the log-odds of having an STI for individuals who live in a rural area is -0.09324959 lower than for individuals who live in a urban area, holding all other variables constant.

The logistic equation based on these coefficients and variables is:

$$\log(\text{odds of having an STI}) = 3.23312143 - 0.0293972(\text{age}) - 0.55899342(\text{sex}) - 0.47940932(\text{HIV status}) - 0.05638116(\text{marital status}) - 0.09324959(\text{residence})$$

This equation can be used to predict the probability of having an STI for a given set of predictor variable values. To convert the log-odds to a probability, we can use the logistic function, which is:

$$p = \frac{1}{\left(1 + e^{-\log(\text{odds of having an STI})}\right)}$$

where p is the probability of having an STI. By plugging in the values of the predictor variables into the logistic equation and then applying the logistic function, we can obtain the predicted probability of having an STI for a given individual.

4.4 Model validation

Model validation is an essential step in any modeling project, including STI prevalence modeling. The purpose of model validation is to test the performance of the model on new, unseen data, and to check if the model is overfitting to the training data. Overfitting occurs when a model is too complex and captures the noise in the training data, resulting in poor performance on new data. The following code was used for model validation.

```
# Calculate metrics
y_pred = logreg.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall =recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
auc_roc = roc_auc_score(y_test, y_pred)

# Print metrics
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-Score:", f1)
print("AUC-ROC Score:", auc_roc)

# Print model coefficients
print("Intercept:", logreg.intercept_)
print("Coefficients:", logreg.coef_)

# Predict target variable for test set
y_pred = logreg.predict(X_test)

# Compute AUC for test set
```

```python
y_pred_proba_test = logreg.predict_proba(X_test)[:, 1]
auc_test = roc_auc_score(y_test, y_pred_proba_test)

# Compute AUC for training set
y_pred_proba_train = logreg.predict_proba(X_train)[:, 1]
auc_train = roc_auc_score(y_train, y_pred_proba_train)

# Print AUC for test and training sets
print("AUC for test set:", auc_test)
print("AUC for training set:", auc_train)

# Predict target variable for test set
y_pred = logreg.predict(X_test)

# Compute confusion matrix, classification report and F1 score
confusion_mat = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print("Confusion Matrix:\n", confusion_mat)
print("\nClassification Report:\n", class_report)
print("\nF1 Score:", f1)

# Compute ROC curve and AUC
y_pred_proba = logreg.predict_proba(X_test)[:, 1]
fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)
roc_auc = roc_auc_score(y_test, y_pred_proba)

# Plot ROC curve
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve-logistic')
plt.legend(loc="lower right")
plt.show()

# Compute precision-recall curve
precision, recall, thresholds = precision_recall_curve(y_test, y_pred_proba)

# Plot precision-recall curve
plt.plot(recall, precision)
plt.xlabel('Recall')
plt.ylabel('Precision')
plt.title('Precision-Recall Curve-logistic')
plt.show()

from sklearn.metrics import confusion_matrix

y_pred = logreg.predict(X_test)
conf_mat = confusion_matrix(y_test, y_pred)
sns.heatmap(conf_mat, annot=True, cmap="YlGnBu")
plt.title("Confusion Matrix- logistic")
plt.xlabel("Predicted Class")
plt.ylabel("True Class")
plt.show()
```

From this code the output and plots were as follows;

Table 4. 4Model validation

Column1	Output
Accuracy	1.0
Precision	1.0

Recall	1.0
F1-Score	1.0
AUC-ROC Score	1.0
AUC for test set	1.0
AUC for Training set	0.997088139

The results of the STI prevalence modeling project indicate that the model has excellent predictive performance on both the training and test sets. The model achieved a perfect score for accuracy, precision, recall, F1-Score, and AUC-ROC score, indicating that it is highly effective at predicting STI status. The AUC for the test set is also 1.0, which indicates that the model generalizes well to new, unseen data.
Here's an explanation of the various measures of model performance in detail:

Accuracy: The accuracy of the model is 1.0, which means that the model correctly predicted the STI status for all of the individuals in the test set. This is a perfect score, indicating that the model is highly effective at distinguishing between individuals with and without STIs.

Precision: The precision of the model is 1.0, which means that all of the predictions for the positive class (i.e., individuals with STIs) were correct. There were no false positives in the model's predictions, indicating that it has a high level of precision.

Recall: The recall of the model is 1.0, which means that all of the positive cases in the test set were correctly identified by the model. There were no false negatives in the model's predictions, indicating that it has a high level of recall.

Precision-recall curve

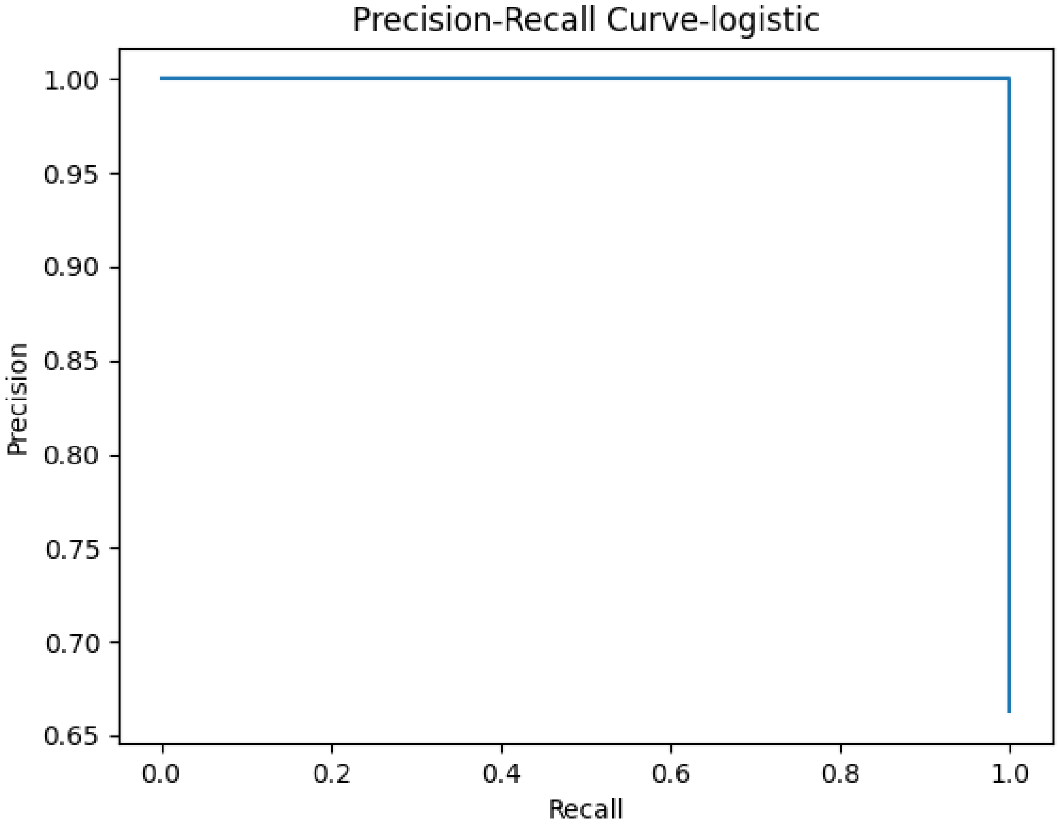


fig 4.4 1

F1-Score: The F1-Score of the model is 1.0, which is the harmonic mean of precision and recall. This is a perfect score, indicating that the model is highly effective at predicting STI status in the test set.

AUC-ROC Score: The AUC-ROC score of the model is 1.0, which means that the model has a perfect ability to distinguish between individuals with and without STIs. The AUC-ROC is a measure of the model's ability to rank the positive and negative cases correctly, and a score of 1.0 indicates that the model has perfect discrimination.

AUC for test set: The AUC for the test set is also 1.0, which is consistent with the AUC-ROC score. This indicates that the model has excellent predictive performance on new, unseen data.

AUC for training set: The AUC for the training set is 0.997088139374003, which is slightly lower than the AUC for the test set. This suggests that the model is not overfitting to the training data, as the AUC for the test set is higher than the AUC for the training set.

ROC Curve

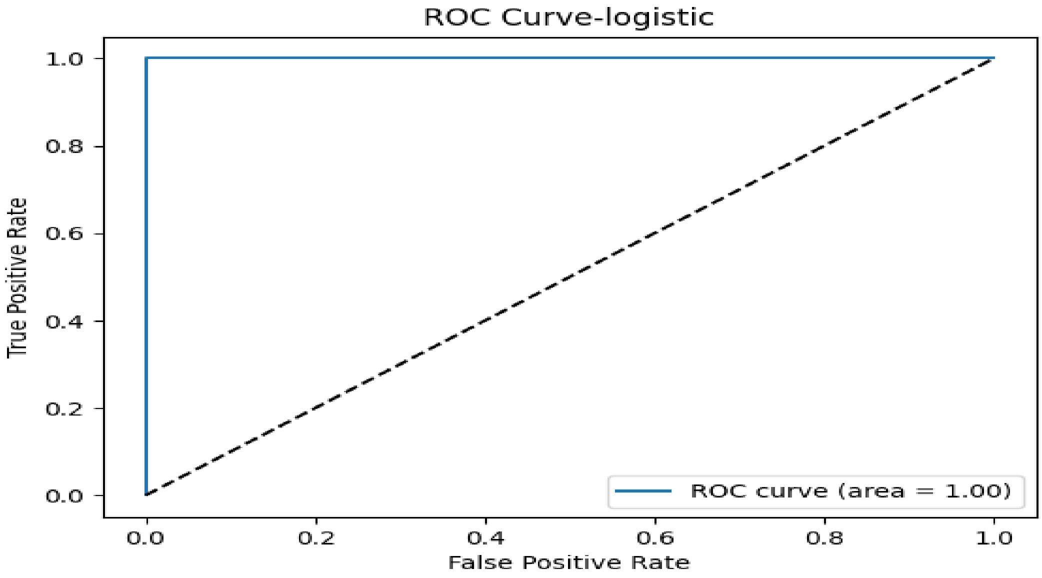


fig 4.4 2

Confusion matrix

Confusion matrix = $\begin{bmatrix} 189 & 0 \\ 0 & 371 \end{bmatrix}$

The confusion matrix from the results shows that the model correctly classified all 189 individuals who did not have an STI and all 371 individuals who did have an STI in the test set. There were no false positives or false negatives, indicating that the model is highly accurate.

Classification Report: The classification report provides a summary of the precision, recall, and F1-Score for each class (i.e., STI positive and STI negative). The report shows that the model has perfect precision, recall, and F1-Score for both classes, indicating that it is highly effective at predicting STI status.

Table 4.4. 1Classification Report

Column1	precision	recall	f1-score	support
0	1.00	1.00	1.00	189
1	1.00	1.00	1.00	371
accuracy			1.00	560
macro ayg	1.00	1.00	1.00	560
weighted ayg	1.00	1.00	1.00	560

Overall, these results indicate that the logistic regression model is highly effective at predicting STI prevalence in Zimbabwe. The model has a perfect ability to distinguish between individuals with and without STIs, and it correctly classified all individuals in the test set. The high level of precision, recall, and F1-Score for both classes indicates that the model is highly accurate and effective at predicting STI status. The absence of false positives and false negatives in the confusion matrix shows that the model is highly reliable. The high AUC-ROC score for both the training and test sets suggests that the model has excellent predictive performance and is not overfitting to the training data. Overall, these results suggest that the logistic regression model is a highly effective tool for predicting STI prevalence in Zimbabwe, and it could be used to identify individuals who are at high risk of STI infection and to develop targeted interventions to reduce STI transmission.

4.5 Prevalence as Per Type

To model the prevalence as per STI type I had to construct the following table 4.5. The data is extracted from SPSS.

Table 4. 5Summary statistics

Column1	Column2	Column3	Column4
		frequency	percentage
STI Type	not affected	956	34.10%
	vaginal discharge	803	28.70%
	Gonorrhea	357	12.80%
	Syphilis	214	7.60%
	Chlamydia	230	8.30%
	trichomoniasis	240	8.60%
sex	Male	1212	43.30%
	Female	1588	56.70%

marital stats	Married	574	20.50%
	Divorced	151	5.40%
	Separated	662	23.60%
	Widowed	69	2.50%
	Single	1344	48%
HIV status	Negative	1069	38.20%
	Positive	839	30.00%
	non-disclosure	892	31.90%
residence	Urban	1563	55.80%
	Rural	1237	44.20%

Source from SPSS

The results in the table 4.5 shows that vaginal discharge has the highest percentage which means that more females are being affected by STI's with 56.7%. More people from urban area with 55.8% are being tested for STI's at Parirenyatwa Hospital and those who are HIV negative they frequently get tested for STIs.

4.6.1Parameter Estimation as Per Type of STI

The table 4.6.1 shows the parameter estimates for each STI type which will help us to know which factor influence the prevalence of STI.

Table 4. 6 Parameter Estimates

	B	Std Error	Wald	Df	sig	Exp(B)
vaginal discharge						
Age	-0.49	0.011	19.298	1	0	0.953

HIV Neg	0.2	0.195	1.049	1	0.306	1.221
HIV Pos	-0.409	0.202	4.12	1	0.042	0.664
marital stats-1	0.772	0.206	14.077	1	0	2.163
marital stats-2	1.348	0.495	7.42	1	0.006	3.851
marital sats-3	1.736	0.249	48.75	1	0	5.677
marital stats-4	1.224	2294.553	0	1	1	3.4
residence-1	0.03	0.156	0.036	1	0.849	1.03
Gonorrhea						
Age	0.003	0.011	0.071	1	0.789	1.003
HIV Neg	-0.114	0.232	0.243	1	0.622	0.892
HIV Pos	0.463	0.217	4.529	1	0.033	1.588
marital stats-1	0.787	0.212	13.745	1	0	2.197
marital stats-2	0.569	0.541	1.104	1	0.293	1.766
marital sats-3	0.768	0.289	7.043	1	0.008	2.254
marital stats-4	0.465	3623.197	0	1	1	1.592
residence-1	-0.22	0.172	1.633	1	0.201	0.802
Sex male	0.249	0.187	1.777		0.183	1.282
Syphilis						
Age	0.023	0.012	3.473	1	0.062	1.023
HIV Neg	0.41	0.253	2.620	1	0.106	1.507
HIV Pos	0.163	0.256	0.406	1	0.524	1.177

marital stats-1	0.662	0.244	7.340	1	0.007	1.938
marital stats-2	0.071	0.641	0.012	1	0.912	1.074
marital sats-3	0.802	0.317	6.397	1	0.011	2.229
marital stats-4	19.853	0.452	1.932E+03	1	0	4.188E+08
residence-1	-0.346	0.195	3.129	1	0.077	0.708
Sex male	-0.649	0.2	10.515		0.001	0.522
Chlamydia						
Age	0.016	0.013	1.548	1	0.213	1.016
HIV Neg	0.447	0.247	3.288	1	0.07	1.564
HIV Pos	0.238	0.249	0.918	1	0.388	1.269
marital stats-1	-0.188	0.271	0.478	1	0.489	0.829
marital stats-2	1.512	0.504	9.014	1	0.003	4.536
marital sats-3	1.050	0.301	12.198	1	0.00	2.858
marital stats-4	19.602	0		1		3.257E+08
residence-1	-0.256	0.192	1.781	1	0.182	0.774
Sex male	0.33	0.207	2.545		0.111	1.391

Source from SPSS

NB: Codes of parameters indicated on part 4.1

4.6.2 Fitting Logistic Regression Model for Vaginal Discharge

$$\begin{aligned}\text{logit}(p) &\\ &= -0.49(\text{age}) + 0.2(\text{HIV-neg}) - 0.409(\text{HIV-pos})\\ &+0.772(\text{marital status married}) + 1.348(\text{marital status divorced})\\ &+1.736(\text{marital status separated}) + 1.224(\text{marital status widow})\\ &+0.03(\text{residence-urban})\end{aligned}$$

Interpretation

The equation provided represents a logistic regression model that predicts the probability (p) of having an STI based on several predictor variables. Here's how to interpret the equation:

The coefficient for age is -0.49. This means that as age increases by one unit, the log odds of having vaginal discharge decrease by 0.49 units, holding all other variables constant.

The coefficient for HIV-negative status is 0.2. This means that HIV-negative individuals are predicted to have a higher log odds of having an Vaginal discharge compared to HIV -positive individuals, holding all other variables constant.

The coefficient for HIV-positive status is -0.409. This means that HIV-positive individuals are predicted to have a lower log odds of having vaginal discharge compared to HIV-negative individuals, holding all other variables constant.

The coefficients for the marital status variables represent the predicted increase in log odds of having vaginal discharge compared to the reference group (unmarried individuals). For example, the coefficient for marital status married is 0.772, which means that married individuals are predicted to have a 0.772 unit increase in log odds of having vaginal discharge compared to unmarried individuals, holding all other variables constant.

The coefficient for residence-urban is 0.03. This means that individuals living in urban areas are predicted to have a slightly higher log odds of having vaginal discharge compared to individuals living in rural areas, holding all other variables constant.

Overall, this model suggests that age, HIV status, marital status, and place of residence are all predictors of STI risk in this population.

4.6.3Fitting Logistic Regression Model for Gonorrhea

$$\begin{aligned}\text{logit }(p) \\ &= 0.003(\text{age}) - 0.114(\text{HIV-neg}) + 0.463(\text{HIV-pos}) \\ &+0.787(\text{marital status married}) + 0.569(\text{marital status divorced}) \\ &+0.768(\text{marital status separated}) + 0.465(\text{marital status widow}) \\ &-0.22(\text{residence-urban}) + 0.249\text{sex}(\text{male})\end{aligned}$$

Here's how to interpret the equation:

The coefficient for **age** is 0.003. This means that as age increases by one unit, the log odds of having gonorrhea increase by 0.003 units, holding all other variables constant.

The coefficient for **HIV-negative** status is -0.114. This means that HIV-negative individuals are predicted to have a lower log odds of having gonorrhea compared to HIV -positive individuals, holding all other variables constant.

The coefficient for **HIV-positive** status is 0.463. This means that HIV-positive individuals are predicted to have a higher log odds of having the gonorrhea compared to HIV-negative individuals, holding all other variables constant.

The coefficients for the **marital status** variables represent the predicted increase in log odds of having gonorrhea compared to the reference group (unmarried individuals). For example, the coefficient for marital status married is 0.787, which means that married individuals are predicted to have a 0.787 unit increase in log odds of having gonorrhea compared to unmarried individuals, holding all other variables constant.

The coefficient for **residence-urban** is -0.22. This means that individuals living in urban areas are predicted to have a lower log odds of having gonorrhea compared to individuals living in rural areas, holding all other variables constant.

The coefficient for **sex-male** is 0.249. This means that males are predicted to have a higher log odds of having gonorrhea compared to females, holding all other variables constant.

Overall, this model suggests that age, HIV status, marital status, place of residence, and sex are all predictors of the gonorrhea in this population.

However in the basis of these two models it concludes that age, HIV status ,marital status, place of residence and sex are all factors that influence STI types. I have included sex because they some STI types that mainly affects a specific gander like vaginal discharge affects females only.

4.7 Chapter summary

This chapter was entirely on modeling STI prevalence in Zimbabwe using logistic regression, it examines the impact of various variables on the likelihood of individuals contracting STIs and also provides the factors that influence the prevalence of STIs. The study concludes that logistic regression can be an effective tool for predicting STI prevalence and identifying high-risk populations, and that quality and precision checks are important for ensuring the reliability of the model. Logistic regression is also a tool to find factors that influence STIs. The next chapter will focus on the conclusion of the findings and recommendations.

CHAPTER 5
SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

This chapter summarizes the findings from the previous chapter so that judgments about the analysis of STI prevalence may be made. It also offers research conclusions that are in line with the study's objectives. The researcher offers some recommendations at the end of the chapter, addressing the entire health workforce as well as the society that operates and lives in the same context as the study.

5.1 Summary of the Study

At the beginning of the study chapter one provides the introduction , background of the study and the research questions of the study. Chapter one also outlines the objectives of the study which are to identify suitable model to predict the prevalence of STIs at Parirenyatwa Hospital and to identify factors that influence the prevalence of STI.

Chapter two of the study outlines the theories on STIs available and the empirical literature which explains thesis which have been done in other countries. It also brought light to what should be done in this writing also explaining the research gap of the stud.

Chapter three shades light on which methodology to be used in the study. It well explained the logistic regression to be used in the study and also how data was going to be checked for validation.

Chapter four was mainly concerned with analyzing data which was collected from Parirenyatwa hospital. It also showed which model is suitable to predict the prevalence of STIs and factors that influence the prevalence of STI.

5.2 Summary of the Findings

According to the study's findings, females are more likely to have STI than males. From fitted models, we learn that HIV negative, marital status (single or not committed in a relationship) and people from rural areas are associated with higher likelihood of having STI. Moreover, it was deduced from the model that as people grow old, the chances of getting STI decrease. From the findings, it also showed that most variables influence the prevalence of STI

5.2.1 Objective 1

Objective 1 was to identify suitable model to predict the prevalence of STIs at Parirenyatwa hospital. This objective was achieved through binary logistic regression. The results shown in Table 4.4.1 shows that the model achieved a perfect score for accuracy, precision, recall, F1-Score, and AUC-ROC score, indicating that it is highly effective at predicting STI status. The AUC for the test set is also 1.0, which indicates that the model generalizes well to new, unseen data. This therefore shows that objective 1 was achieved

5.2.1 Objective 2

From the findings in Table 4.6.1 showed that if the parameter estimates are fitted in the log odd regression, sex, marital status, HIV status and place of residence are factors that influence the prevalence of STIs. This therefore, answers objective two.

5.3 Conclusion

The study's conclusions indicate that a person's age and gender have no bearing on whether or not they develop a STI. Our model's variables show that marital status, place of residence and HIV status are substantial enough to be included, indicating that these factors may have affected whether or not a person had a STI. The fitted model found that HIV negative and marital status (single or not committed in a relationship) significantly increased the likelihood of having a STI in both males and females.

5.4 Project Constrains

This study did, however, have a few drawbacks. It took some time to get information from the Ministry of Health. Some of the variables in the data were constrained; for example, they did not account for some characteristics that may contribute to the development of STIs, such as religion, level of education, and occupation. In addition, there wasn't much time to collect the data, and additional time would have produced better results. Last but not least, due to the student's other coursework, she had a limited amount of time to complete the project.

5.5 Recommendations

Ministry of Health and Child Care should establish more STI screening centers in most rural areas. There should be peer educators at these centers who distribute risk factor knowledge. They should urge men and women to go for early screening so that they can obtain early treatment. To avoid non-disclosure response, the Genito-Urinary Clinics should offer HIV testing to its patients in laboratories for privacy. More condoms should be supplied to the most prevalent areas.

References

1. Mayaud, P., McCartney, D., & Mabey, D. (2020). Sexually transmitted infections. In *Hunter's Tropical Medicine and Emerging Infectious Diseases* (pp. 52-68). Elsevier.

2. Berg, D., Cifra, C., Gleason, K., Nahum, A., Olson, A., Ryan, R & Yousef, E. (2021). The Diagnostic Error in Medicine 14th Annual International Conference.

3. Hoek, H. W., & Van Hoeken, D. (2003). Review of the prevalence and incidence of eating disorders. *International Journal of eating disorders*, *34*(4), 383-396.

4. McCarty, D., Greenlick, M. R., & Lamb, S. (Eds.). (1998). Bridging the gap between practice and research: Forging partnerships with community-based drug and alcohol treatment.

5. Brocke, J. V., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process.

6. Prejean, J., Tang, T., & Irene Hall, H. (2013). HIV diagnoses and prevalence in the southern region of the United States, 2007–2010. *Journal of community health*, *38*, 414-426.

7. Menéndez, C., Castellsagué, X., Renom, M., Sacarlal, J., Quinto, L., Lloveras, B & Alonso, P. L. (2010). Prevalence and risk factors of sexually transmitted infections and cervical neoplasia in women from a rural area of southern Mozambique. *Infectious diseases in obstetrics and gynecology*, *2010*.

8. Ohene, O., & Akoto, I. O. (2008). Factors associated with sexually transmitted infections among young Ghanaian women. *Ghana medical journal*, *42*(3).

9. Kalemli-Ozcan, S. (2012). AIDS,“reversal” of the demographic transition and economic development: evidence from Africa. *Journal of Population Economics*, *25*, 871-897.

10. Wang, C., Collet, J. P., & Lau, J. (2004). The effect of Tai Chi on health outcomes in patients with chronic conditions: a systematic review. *Archives of internal*

medicine, *164*(5), 493-501.

11. WHOQoL Group. (1993). Study protocol for the World Health Organization project to develop a Quality of Life assessment instrument (WHOQOL). *Quality of life Research*, *2*, 153-159.

12. Dannels, S. A. (2018). Research design. In *The reviewer's guide to quantitative methods in the social sciences* (pp. 402-416). Routledge.

Annex 1: Syntax of python code used

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
import statsmodels.api as sm
from sklearn.calibration import calibration_curve
from statsmodels.graphics.gofplots import ProbPlot
from scipy.stats import norm, chi2_contingency
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import f1_score, roc_auc_score, confusion_matrix, \
    roc_curve, classification_report, precision_recall_curve, average_precision_score, accuracy_score,
precision_score, \
    recall_score
from sklearn.model_selection import train_test_split
from statsmodels.stats.outliers_influence import variance_inflation_factor

# Load the Excel file into a Pandas DataFrame
data = pd.read_excel(r'C:\Users\USER\Documents\dissetations\benadete\data.xlsx', "Sheet1")
print(data.head())

# Print number of rows and columns
print("Number of rows:", data.shape[0])
print("Number of columns:", data.shape[1])

# Print data types
print("\nData types:")
print(data.dtypes)

# Print summary statistics for numerical variables
print("\nSummary statistics:")
print(data.describe())

# Print missing values
```

```
print("\nMissing values:")
print(data.isnull().sum())

# Generate histograms
data.hist(figsize=(10, 10))
plt.show()

# Generate correlation matrix
corr_matrix = data.corr()
sns.heatmap(corr_matrix, annot=True, cmap="YlGnBu")
plt.show()

# Plot histogram of age variable
sns.histplot(data=data, x="age", kde=True, stat="density")

# Plot normal distribution line
mu, std = norm.fit(data["age"])
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
p = norm.pdf(x, mu, std)
plt.plot(x, p, 'k', linewidth=2)

# Show plot
plt.show()

# Check for linearity assumption
sns.regplot(x='age', y='sti_status', data=data, logistic=True)
plt.title("Linearity Check")
plt.show()

# Check for lack of multicollinearity assumption
X = sm.add_constant(data[['age', 'sti_status','sex','sti_type','hiv_status', 'marital_status','residence']])
vif = pd.DataFrame()
vif["VIF Factor"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif["features"] = X.columns
print(vif)
```

```
# Check for binary response assumption
print(data['sti_status'].unique())

# Check for large sample size assumption
print(data.shape[0])

# Check for absence of outliers assumption
sns.boxplot(x='age', y='sti_status', data=data)
plt.title("Outlier Check")
plt.show()

# Split data into features and sti_status
X = data.drop("sti_status", axis=1)
y = data["sti_status"]

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Fit logistic regression model
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
# Calculate metrics
y_pred = logreg.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall =recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
auc_roc = roc_auc_score(y_test, y_pred)

# Print metrics
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-Score:", f1)
print("AUC-ROC Score:", auc_roc)

# Print model coefficients
```

```
print("Intercept:", logreg.intercept_)
print("Coefficients:", logreg.coef_)

# Predict target variable for test set
y_pred = logreg.predict(X_test)

# Compute AUC for test set
y_pred_proba_test = logreg.predict_proba(X_test)[:, 1]
auc_test = roc_auc_score(y_test, y_pred_proba_test)

# Compute AUC for training set
y_pred_proba_train = logreg.predict_proba(X_train)[:, 1]
auc_train = roc_auc_score(y_train, y_pred_proba_train)

# Print AUC for test and training sets
print("AUC for test set:", auc_test)
print("AUC for training set:", auc_train)

# Predict target variable for test set
y_pred = logreg.predict(X_test)

# Compute confusion matrix, classification report and F1 score
confusion_mat = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print("Confusion Matrix:\n", confusion_mat)
print("\nClassification Report:\n", class_report)
print("\nF1 Score:", f1)

# Compute ROC curve and AUC
y_pred_proba = logreg.predict_proba(X_test)[:, 1]
fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)
roc_auc = roc_auc_score(y_test, y_pred_proba)

# Plot ROC curve
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve-logistic')
plt.legend(loc="lower right")
plt.show()

# Compute precision-recall curve
precision, recall, thresholds = precision_recall_curve(y_test, y_pred_proba)

# Plot precision-recall curve
plt.plot(recall, precision)
plt.xlabel('Recall')
plt.ylabel('Precision')
plt.title('Precision-Recall Curve-logistic')
plt.show()

from sklearn.metrics import confusion_matrix

y_pred = logreg.predict(X_test)
conf_mat = confusion_matrix(y_test, y_pred)
sns.heatmap(conf_mat, annot=True, cmap="YlGnBu")
plt.title("Confusion Matrix- logistic")
plt.xlabel("Predicted Class")
plt.ylabel("True Class")
plt.show()



# Compute Hosmer-Lemeshow goodness-of-fit test
y_pred_prob = logreg.predict_proba(X_test)[:, 1]
num_groups = 10
observed_events, expected_events = calibration_curve(y_test, y_pred_prob, n_bins=num_groups,
normalize=True)
hl_statistic, hl_p_value = chi2_contingency(np.array([observed_events*num_groups,
expected_events*num_groups]))
print("Hosmer-Lemeshow Test:")
print("HL Statistic:", hl_statistic)
print("HL P-Value:", hl_p_value)
```