FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 5 No. 1, March, 2021, pp 445 – 456
DOI: https://doi.org/10.33003/fjs-2021-0501-589

ASSESSING THE PERFORMANCE OF SOME RE-SAMPLING METHODS USING LOGISTIC REGRESSION

[1]Usman, U., [1]Waziri, M., [*1]Manu, F., [2]Zakari, Y. & [2]Dikko, H. G.

[1]Department of Mathematics, Usmanu Danfodiyo University, Sokoto, Nigeria.
[2]Department of Statistics, Ahmadu Bello University, Zaria, Nigeria.

[*]Corresponding author's email: farukmanu9@gmail.com

ABSTRACT
This research reports on the performance of two re-sampling methods (Bootstrap and Jackknife) relationship and significance of social-economic factors (age, gender, marital status and settlement) and modes of HIV/AIDS transmission to the HIV/AIDS spread. Logistic regression model, a form of probabilistic function for binary response was used to relate social-economic factors (age, sex, marital status and settlement) to HIV/AIDS spread. The statistical predictive model was used to project the likelihood response of HIV/AIDS spread with a larger population using 10,000 Bootstrap re-sampled observations and Jackknife re-sampled. From the analysis obtained from the two re-sampling methods, we can conclude that HIV transmission in Kebbi state is higher among the married couples than single individuals and concentrate more in the rural areas.

Keywords: Re-sampling, Logistic regression, Odd ratio, HIV and Category data

INTRODUCTION

Background to the study

Re-sampling is a method for estimating the precision of sample statistics (medians, variances, percentiles) by using subsets of available data (jackknifing) or drawing randomly with replacement from a set of data points (bootstrapping). It is also a method for exchanging labels on data points when performing significance tests (permutation tests, also called exact tests, randomization tests, or re-randomization tests) and for validating models by using random subsets (bootstrapping, cross validation).

Re-sampling does not emerge without any context. Indeed, the resampling method is tied to the Monte Carlo simulation, in which researchers "make up" data and draw conclusions based on many possible scenarios.

Methods of re-sampling are playing a vital role in statistical inference, their applications become more popular especially in the last 2 decades. Researchers (such as Bello *et al* (2015a), Bello *et al* (2015b), Bobbit*et al.*, (2007) and Wu*et al*(1986)applieda class of weighted jackknife variance estimators for the least squares estimator by deleting any fixed number of observations at a timeand arrived at a conclusion. Bello *et al.*, used only Bootstrap Re-Sampling method to a categorical data of HIV/AIDS and considered only age, sex, and employment statuswhile marital status and Rural/Urban settlements are significant factors of HIV/AIDS transmission and Bobbit*et al.*, also used Bootstrap Application method in the International Price Program. In this research, we will study only two re-sampling methods (Bootstrap and jackknife) simultaneously in the same phenomena so as to see the performance of each and use logistic regression to finally obtain the best method among them.

Bootstrap: This technique was invented by Efron (1979, 1981, 1982) and further developed by Efron and Tibshirani (1993). "Bootstrap" means that one available sample gives rise to many others by re-sampling (a concept reminiscent of pulling yourself up by your own bootstrap).

The general variance estimation procedures is based on the following substitution idea. The functional form of $\sigma^2\ (\theta)$ (as a function of θ). It is natural to simply estimates $\sigma^2\ (\theta)$ by replacing the unknown parameter θ and $\hat{\theta}$, namely use as variance estimate $\widehat{\sigma^2} = \sigma^2\left(\hat{\theta}\right)$, whether σ^2 itself is a reasonable estimate of $\sigma^2\left(\hat{\theta}\right)$ or not. In order for this procedure to be reasonable $\sigma^2\left(\theta\right)$

need to be a continuous function of θ, and $\hat{\theta}$ would have to be reasonable estimate of θ ,i.e $\hat{\theta}$ be sufficiently near θ. The applicability of the above natural substitution procedure is quite general and it can be carried out provided we have the functional form of $\sigma^2\ (\theta)$ as a function of θ. Bootstrap method provides a simple algorithm to for getting accurate approximations to σ^2.

Jackknife: This was introduced by Quenouille 1949. Quenouille's aim was to improve an estimate by correcting for its bias. Later on Tukey(1958)popularized the method and found that a more important use of the jackknife was to estimate standard errors of an estimate. The name "Jackknife" was coined by Tukey to imply that is astatistical tool with many purpose. Jackknife is a step beyond cross-validation. In Jackknife, the same test is repeated by leaving one subject out each time. Thus, this technique is also called leave one out. The bootstraping uses the bootstrap samples to estimate variability, while jackknife uses what are called pseudovalues.
sample mean.

HIV/AIDS and its mode of transmission

A Human Immunodeficiency Virus known as HIV, attacks the body's Immune system, specifically the CD4 cells (T cells), which helps the immune system fight off infections.
There are four main routes of HIV transmission:

1. Unprotected Vaginal or anal or oral sex: HIV can be transmitted from woman to man or vice versa. Open cuts and sores increase the risk. (Oral sex has a very small risk, but only if there are sores in/around the mouth or on the receiving partner's genitals.)
2. Injecting drugs: Shared un-sterilized equipment can carry infected blood. Needles used for body piercing and tattooing can carry a small risk.
3. Blood transfusions/transplants: All donated blood should be tested for HIV; any untested blood carries a risk of transmission.
4. Mother-to-child: Transmission can occur during pregnancy, labour, delivery or breastfeeding if treatment is not taken correctly.

LITERATURE REVIEW

Introduction

Bootstrap re-sampling method researches began at the end of the 70s decade, although a lot of related projects can be verified before that period. The most theoretical development was elaborated after 1979, with Efron(1983). We have several motivations to use these methods, for example when usual model assumptions can't be verified in certain dataset. For example non normal distribution with outliers or mixed distribution with errors. We can also point to asymptotic results when the number of available observations aren't enough to guarantee the asymptotic convergence. In these cases, methods based on simulation, more specifically re-sampling, can be useful to establish the uncertainty about estimation of the parameters.

The logistic regression model is one of the popular statistical models for the analysis of binary data with applications in physical, biomedical, behavioral sciences, and many others. Logistic regression analysis was implemented to determine the significant contributory factors influencing the subject of study. The cases having the response variable as categorical, often called binary of (yes/no; present/absent; etc) and possible explanatory variables which can either be categorical variables, numerical variables or both are numerous in the biometry, psychometric, and epidemiology researches.

Empirical Evidence.

Bello *et al.,* (2015b), Bobbit*et al.,* (2007) and Wu *et al.,* (1986) applied a class of weighted jackknife variance estimators for the least squares estimator by deleting any fixed number of observations at a time and arrived at a conclusion. Bello *et al.,* used only Bootstrap Re-Sampling method to a categorical data of HIV/AIDS and Bobbit*et al.,* also used Bootstrap Application method in the International Price Program. Both of them uses bootstrap re-sampling only meanwhile other re-sampling methods can be tested and found to be essential in those aspect.

L.M Raposo*et al.*,(2013) used the logistic regression model to predict resistance to HIV protease Inhibitor, the model obtained was said to be useful in decision making regarding the best therapy for HIV positive individuals.

The work of J. Ren*et al.*,(2014), was suspicious of the suitability of ordinary, categorical exposures, and logarithm transformation functions presented in logistic regression model to assess if the likelihood of infectious diseases is risk or as a result of exposure using simulated data, the work adjudged the logarithmic transformation function as better than the other two.

However, they found that risk of using logistic regression is not a risk at all if large sample size is used or procedure of large sample technique such as bootstrap re-sampling method is used. This reduced the bias in their estimates. "The odd function is the most suitable function for interpretation of binary predictive problems".

Bello *et al.,* (2015), investigate on the relationship and significance of social-economic factors (age, gender, employment status) and modes of HIV/AIDS transmission to

the HIV/AIDS spread in Oyo state, Nigeria. They used logistic regression model, a form of probabilistic function for binary response was used to relate social-economic factors (age, sex, employment status) to HIV/AIDS spread. They also used statistical predictive model to project the likelihood response of HIV/AIDS spread with a larger population using 10,000 Bootstrap re-sampled observations. In their findings the age group as block effect shows adequate significant level to HIV infection with F-value 6.496 and significant level 0.004($p<0.05$). The age group 16-39 seems to be the age block that is most infected in the population, this age group is the reproductive age and the most sexually active stage of any population which suggests that any additional to the uncontrolled activities of sexual intercourse and pregnancy without proper medical supports will increase the cases of mother-to- child infection in particular. An individual will not contract HIV because he/she belongs to a particular gender; contraction is majorly as result of activities or exposure. They recommended increment in employment level especially for the female gender in Oyo state as a vital control measure to mitigate the spread of HIV/AIDS coupled with increase in public awareness, abstinence, and a more comprehensive approach to preventing mother-to-child infection. The researchers did not consider marital status age group and rural /urban communities as a social economic factors which are also significant in the modes of transmission of HIV/AIDS. Apart from Bootstrap other re-sampling methods can be used to make a good estimation.

RESEARCH METHODOLOGY

Introduction

Considering the case where our response y_i is a dichotomous response, when possible response is either yes or no, death or alive, present or absent and as the case at hand in this work is either HIV negative or HIV positive.

$$y_i = \begin{cases} 1 & if\ ith \text{ individual is HIV positive} \\ 0 & if\ ith \text{ individual is HIV negative} \end{cases}$$

We coded the present or absent of subject of study as 1 and 0 respectively. The distribution of y_i is binomial of a single trial or basically Bernoulli distribution as used by some text. The binary indicator variable outcome can only be 1 or 0 as the probability is bound between 0 and 1; this gives a sigmodial shape approaching 0 and 1 asymptotically. This is a nonlinear problem.

Logistic Regression: The logistic function relating $y_i$to predictors which can be qualitative, quantitative or both is a very flexible model which makes it vital to solving many epidemiology and social indicators related problems. The logistic line can also be of form;

$$y_i = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k \qquad i = 0,1 \qquad (3.1)$$

$$E(y) = E(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k)) \qquad (3.2)$$

Probabilities; given $P(y_i = 1) = \pi_i \text{and} P(y_i = 0) = \pi_i$,

therefore, $E(y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi$ (3.3)

π is the probability of our subject of interest in study taking place and 1- π is the probability of subject ofinterest not occurring. The subject of interest informs our choice of coding 1 or 0. Equation (3.3) gives the probability y_i given that level of parameter variable isx_1. Logistic regression model is a special case of general linear model. The special problems associated with model having binary response variable is the problem of having our error terms not normally distributed and heteroskadastic in nature due to the distribution of our response variable bonded between 0 and 1.

$$e_{i\,=}\ y_i - (\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k)$$

$$y_i = 1; \varepsilon = 1 - \theta_0 - \theta_1 x_1 - \theta_2 x_2 - \ldots - \theta_k x_k \qquad (3.3.1)$$

$$y_i = 0; \varepsilon = -\theta_0 - \theta_1 x_1 - \theta_2 x_2 - \ldots - \theta_k \text{xk}$$

From equation (3.3.1) ε is not normally distributed. Other problem associated with the logistic model is the constraint condition on response function. The function form in equation (3.3) has its left hand-side to take value ranging between 0 and 1, while the right hand-side is not in a form that can return values between 0 and 1 asymptotically. Therefore, we require a link function to properly link the left hand-side to the right hand-side. Link function such as identity function will not be appropriate for the nonlinear problem at hand. However, out of several possible link functions, we shall use the logit function, for easier understanding and interpretations of our results.

The model is written in the form;

$$\text{logit}(\pi) = \log\frac{\pi}{1-\pi} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k \qquad (3.4)$$

The odds can vary on a scale of $(0,\infty)$, so the log odds can vary on the scale of $(-\infty,\infty)$ – precisely what we get from the right hand side of the linear model. For a real-valued explanatory variable x_i, the intuition here is that a unit additive change in the value of the variable should change the odds by a constant multiplicative amount.

Exponentiating, this is equivalent to equation 3.4

$$e^{\text{logitp}} = e^{\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k} \qquad (3.5)$$

The interpretation of θ's is not straightforward because the increase in unit of X varies for the logistic regression model according to the location of the starting point of the X. The logit function is the natural logarithm (ln) of odds of y and taking exponential of the log of odd function gives us the most appreciable odd function that is vital in our interpretation of result. The odd function will simplify the interpretation problem.

$$\frac{\pi}{1-\pi} = e^{\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k} \qquad (3.6)$$

Explicitly;

$$\pi = \frac{e^{\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k}}{(1+e^{\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k})} \qquad (3.7)$$

The inverse of the logit function is the logistic function.

Hence;

$$\pi = probability(0,1/X=x) = \frac{e^{\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k}}{(1+e^{\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k})} \qquad (3.8)$$

The logistic function form in equation (3.7) and (3.8) will return the right hand-side to be property value ranging from 0 and 1. The function increases monotonically if the gradient $\theta>0$and decreases monotonically if $\theta < 0$.

Method of Estimation

The variability of the error terms variances differs at different level of X, as shown in equation (3.3.1). This makes Ordinary Least square estimation ineffective in estimation of logistic function. The maximum likelihood is a better method for estimating logistic function since logistic function predicts probabilities, and not just classes, it can fit the probabilities for each class of our data-point, either for the class 'π_i ' or '1-π_i'. We must also note that the error term is not usually considered in logistic problems.

Maximum Likelihood Estimation

The maximum likelihood estimate is that value of the parameter that makes the observed data most likely (3.12). The values of θ s that maximize $\log_e L(\theta)$, that is, the value of θ that assign the highest possible probability to the sample that was actually obtained. The method of likelihood in estimating a logistic function usually requires numerical procedures, and Fisher scoring or Newton-Raphson often work best. Most statistical packages have the logit numerical search procedure. In this work, R-programming language package for obtaining the maximum likelihood estimates of a logistic regression is used.

Let $y_1, y_2, \ldots y_k$ be n independent random variables (r.v.'s) with probability density functions $f(y;\theta)$ that depends on parameter θ. The likelihood of the joint density function of k independent observations is $L(\theta) = f(y_1, y_2, \ldots y_k, \theta)$

or $$f(y_1, y_2, \ldots, y_k / \theta) = f(y_i / \theta) \ldots f(y_k / \theta) \quad (3.9)$$

then $$f(y;\theta) = \prod_{i-1}^{n} f_i(y_i;\theta) = L(\theta; y) \quad (3.10)$$

The root of the equation is obtained by equating the first derivative of equation (3.2.2) to zero and solved for each parameter. The maximum likelihood estimate (MLE) hold only when the second derivative is negative. The probability distribution function of our y_i follows the Bernoulli distribution, u with y_i taking zero or one. The likelihoodfunction is;

$$L(\theta_i) = \prod_{i=1}^{n} \pi_{(x_i)}{}^{y_i} (1-\pi_{(x_i)})^{1-y_i} = \prod_{i=1}^{n} (\pi_{(x_i)})^{y_i} (1-\pi_{(x_i)})^{-y_i} (1-\pi_{(x_i)})^{1}$$

$$\prod_{i=1}^{n} \left(\frac{\pi_{(x_i)}}{1-\pi_{(x_i)}} \right)^{y_i} (1-\pi_{(x_i)}) \quad (3.11)$$

Recall equation (3.7), π is substituted into equation (3.11)

$$= \prod_{i=1}^{n} \frac{(e^{\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k})^{y_i}}{1+e^{\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k}} \quad (3.12)$$

taking the natural logarithm or the log-likelihood function yields,

$$l(\theta) = \sum_{i=1}^{n} y_i \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k - \sum_{i=1}^{n} In(1+e^{\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k}) \quad (3.13)$$

the first derivative of the log-likelihood function gives the gradient.

The first derivative of $\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k$ with respect to $\theta_j is x_{ij}$ so

$$\frac{\partial l(\theta)}{\partial \theta_j} = \sum_{n=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{e^{\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k}}{1+e^{\theta_0+\theta_1 x_1+\theta_2 x_2+\ldots+\theta_k x_k}} x_{ij}$$

Recall equation (3.1.7) and substitute forπ; the probability of subject of interest under study occurring

$$\frac{\partial l(\theta)}{\partial \theta_j} = \sum_{n=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \pi_i x_{ij} \qquad j = 1, 2, ..., k \tag{3.14}$$

$$= \sum_{n=1}^{n} (y_i - \mu_i) x_{ij} \qquad \text{where} \quad \mu_i = E(y_i) = \pi_i$$

The differentiation of the log likelihood function in equation (3.13) with respect to each parameter θ_j will not analytical give us the maximum likelihood estimates by setting each of the k equations in equation (3.13) equal to zero. It is a system of k nonlinear equations. The solution to the K unknown variables is a nonlinear problem cannot be solved analytically but through numerical estimation using an iterative process. The Newton-Raphson method is popularly used for a logistic nonlinear function. However, problem of multi-collinarity may arise which is visible when there are large estimated parameters and large standard error values. Also, convergence problem in numerical search procedure can be associated with multi-collinearity problem which can be overcome by reducing the number of parameter variables for easy and quick convergence.

Variance Estimation of a Logistic Function Using the 4 Re-sampling Method

The general linear model rely on asymptotic approximations in estimating the coefficient standard errors and this may not be reliable, just as measures such as R-square based, residual errors are not very informative and can be misleading. Therefore, using thesemethodsa re-sampling technique (Bootstrap, Jackknife, Randomization exact test and Cross validation) will either confirm or dispel our doubts about the sufficiency of our sample to estimate unbiased and robust estimates for the population parameters. For our models to adequately capture the reality of HIV/AIDS spread across different socio-economical classes in Kebbistatepopulation as likely as possible. We shall Generate 10,000 Re-sampling samples from the original sample to estimate our models' parameter values and their confidence intervals**.**

The Odd Function

Bland and Douglas (2000) mentioned that there are mainly three reasons to use the odds ratio. "Firstly, they provide an estimate (with confidence interval) for the relationship between two binary variables. Secondly, they enable us to examine the effects of other variables on that relationship, using logistic regression. Thirdly, they have a special and very convenient interpretation." The odds are nonnegative, with odds 1.0 when a success is more likely than a failure. According to Pedhazur (1997) Odds are determined from probabilities and range between 0 and infinity. Odds are defined as the ratio of the probability of success and the probability of failure. The odds of success given as π/1-π and the odds of failure would be odds (failure) given as $1-\pi/\pi$. The odds of success and the odds of failure are just reciprocals of one another. Probability and odds both measure how likely it is that our subject of interest will occur. Notably, the sign of the log-odds ratio indicates the direction of its relationship, the distinction regarding a positive or negative relationship in that of the odds ratios is given by which side of 1 the odd values fall on. Odd value 1 indicates no relationship, less than one indicates a negative relationship and greater than one indicates a positive relationship. However, in order to get an intuitive sense of how much things are changing, we need to get the exponential of the log-odds ratio, which gives us the odds ratio itself.

The odd ratio of the odd for x=1 to the odd of x=0 is

$$\text{the odd ratio} = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)} \tag{3.15}$$

$$= \left(\frac{\dfrac{e^{\theta_0+\theta_1+\theta_2+...+\theta_k}}{(1+e^{\theta_0+\theta_1+\theta_2+...+\theta_k})}}{\dfrac{1}{(1+e^{\theta_0+\theta_1+\theta_2+...+\theta_k})}} \right) \div \left(\frac{\dfrac{e^{\theta_0}}{(1+e^{\theta_0})}}{\dfrac{1}{(1+e^{\theta_0})}} \right)$$

$$= \frac{e^{\theta_0+\theta_1+\theta_2+\ldots+\theta_k}}{e^{\theta_0}}$$

The odd ration = $\exp(\theta_1)*\exp(\theta_2)*\ldots*\exp(\theta_k)$ (3.16)

This result obtained is the relationship between the odds ratio and an independent dichotomous. The result tells that the odds on the event that y equals 1, increases (or decreases) by the factor $\exp(\theta_1+\theta_2+\ldots+\theta_k)$ among those with x= 1 than among those x= 0. One major condition to note when interpreting for multiple logistic regression is that the estimated odds ratio for predictor variable x assumes that all other predictor variables are held constant.

RESULTS AND DISCUSSION

BOOTSTRAP RESULTS

Table 1. Bootstrap Estimates

Parameters	Original Estimates	Bootstrap Estimates	Bootstrap Standard Error	Bootstrap Confidence interval	
				2.5%	97.5%
Intercept	-2.207841	-2.207841	0.624581	-3.478559	-1.01893
Age	0.006179	0.006179	0.014885	-0.02392	0.03469
Sex	-1.118444	-1.118444	0.318448	-1.76733	-0.51133
Marital Status	0.417884	0.417884	0.412343	-0.34169	1.29414
Settlement	0.041535	0.041535	0.321572	-0.58235	0.68404

The residuals plotted against the predicted probability (See figure 2), shows the lowess smooth approximates a line having zero slope and intercept, and we can conclude that model inadequacy is not apparent.

The parameter estimates from the original observation and the 1,000 bootstrap samples were asymptotically the same and both confidence intervals coincide at almost the same intervals; these demonstrate the precision of the model coefficient estimates. Thus, we can conclude within approximately 95 percent confidence that our sample size is as sufficient as using any other larger sample size, (See table 1). Statistically, we can say that our sample size is a good representation of the entire population from which it was drawn.

Recall

$$H_i = prob(0,1/\ X = x) = \left[1+\exp\left(-\theta_0-\theta_1 x_1-\theta_2 x_2-\ldots-\theta_k x_k\right)\right]^{-1} \quad (4.1)$$

$$H = \left[1+\exp\left(2.207841-0.006179Age+1.118444Sex-0.417884Settl-0.041535Msta\right)\right]^{-1} \quad (4.2)$$

Table 2. Multiple Sample parameters Estimates

Parameters	Bootstrap Estimates 1,000	Original Estimates 500 sample	Original Estimates 400 sample	Original Estimates 300 sample	Original Estimates 200 sample	Original Estimates 100 sample
Intercept	-2.207841	-2.207841	-2.278947	-1.57388	-1.714911	-0.93678
Age	0.006179	0.006179	0.006854	-0.01204	-0.004955	-0.04497
Sex	-1.118444	-1.118444	-1.323870	-1.37729	-1.263447	-1.65770
Marital Status	0.417884	0.417884	0.149441	0.07102	0.194693	0.73991
Settlement	0.041535	0.041535	0.695419	0.55746	0.550365	0.18569

Model for 100 samples

$$H = \left[1+\exp\left(2.207841-0.006179Age+1.118444Sex-0.417884Settl-0.041535Msta\right)\right]^{-1}$$

The equation (4.2) is from the 500 samples. Equation 4.3 was fitted from different samples of 100, 200, 300 and 400 data points respectively.

The positive coefficient value of Age parameter suggests a positive relationship between age and HIV infection, which imply that the probability of contracting HIV increases as Age of person(s) increases. The odds of contracting HIV given age cannot be given a direct interpretation based on the question 'what unit of age is appropriate and applicable to show the change in odds ratio?' The odds is best described by exp(c*Age), given c is a difference of units of ages under comparison. For the difference of unit between age 25 and 35years, the odds of contracting HIV between age 25 and 35years is exp (10*(0.006179)= 1.063739. which indicate a positive relationship between age and HIV infection. The relationship between age and Probability of HIV infection suggests that the older generation above 35years should be of first priority in all the agenda towards eradicating HIV/AIDS spread.

The negative coefficient value of sex parameter suggests a negative relationship between age and HIV infection even though it shows great statistical significance at 0.05 level of significance with standard error of 0.318448. However, an individual will not contact HIV because he/she belongs to particular sex, and contraction is majorly as a result of activities or exposure.

Also from the table 1 above, the difference in the odds ratio of HIV infection between the male married individuals in the urban population and male single individuals in the urban population is 0.1767087. This result implies that the odds of male married individuals in the urban population contracting HIV are 18% higher than that of the male single individuals in the urban population for given age. The difference in the odds ratio of HIV infection between the male married individuals in the rural population and their counterpart is 0.1695194. This result implies that the odds of male married individuals in the rural population contracting HIV are 17% higher than that of the male single individuals for given age.

The odds ratio of HIV infection between female married individuals in urban population is 54% higher than those unmarried female individuals in the same population. Likewise an odd of HIV infection between married female individuals in rural population is 52% higher than that of unmarried counterpart.

PREDICTIVE MODELS

We can now predict the likelihood of HIV spread in kebbi State among different sex, marital status and settlement across all possible age.

1. The predicted model for male marries individuals in urban population.

$$\text{Prob.oddsMMU} = \frac{\exp\left(-2.207841+0.006179*Age-1.118444+0.417884+0.041535\right)}{\left(1+\exp(-2.207841+0.006179*Age-1.118444+0.417884+0.041535)\right)} \quad (4.3)$$

2. The predicted model for male single individuals in urban population.

$$\text{Prob.odds MSU} = \frac{\exp\left(-2.207841+0.006179*Age-1.118444+0.041535\right)}{\left(1+\exp(-2.207841+0.006179*Age-1.118444+0.041535)\right)} \quad (4.4)$$

3. The predicted model for male marries individuals in rural population.

$$\text{Prob.odds.MMR} = \frac{\exp\left(-2.207841+0.006179*Age-1.118444+0417884\right)}{\left(1+\exp(-2.207841+0.006179*Age-1.118444+0.417884)\right)} \quad (4.5)$$

4. The predicted model for male single individuals in rural population

$$\text{Prob.odds.MSR} = \frac{\exp\left(-2.207841+0.006179*Age-1.118444\right)}{\left(1+\exp(-2.207841+0.006179*Age-1.118444)\right)} \quad (4.6)$$

5. The predicted model for female marries individuals in urban population

$$\text{Prob.odds.FMU} = \frac{\exp\left(-2.207841+0.006179*Age+0417884+0.041535\right)}{\left(1+\exp(-2.207841+0.006179*Age+0.417884+0.041535)\right)} \quad (4.7)$$

6. The predicted model for female single individuals in urban population

$$\text{Prob.odds.FSU} = \frac{\exp\left(-2.207841+0.006179*Age+0.041535\right)}{\left(1+\exp(-2.207841+0.006179*Age+0.041535)\right)} \quad (4.8)$$

7. The predicted model for female marries individuals in rural population

$$\text{Prob.odds FMR} = \frac{\exp(-2.207841 + 0.006179 * Age + 0.417884)}{(1 + \exp(-2.207841 + 0.006179 * Age + 0.417884))} \quad (4.9)$$

8. The predicted model for female single individuals in rural population

$$\text{Prob.odds.FSR} = \frac{\exp(-2.207841 + 0.006179 * Age)}{(1 + \exp(-2.207841 + 0.006179 * Age))} \quad (4.10)$$

JACKNIFE RESULTS

Table 2. Jackknife Estimate

Parameters	Jackknife Estimates	Jackknife Confidence Interval		Jackknife Standard Error
		2.5%	97.5%	
Intercept	-2.33742	-3.06895	-1.64502	0.36192
Age	-0.34753	-1.92933	0.97540	0.73707
Sex	0.24441	-0.05570	0.55137	0.15411
Marital Status	0.18739	-0.26219	0.60035	0.21841
Settlement	-0.01356	-0.37788	0.33729	0.18149

The negative coefficient value of Age parameter suggests a negative relationship between age and HIV infection, which imply that the probability of contracting HIV decreases as Age of person(s) increases. The odds is best described by exp(c*Age), given c is a difference of units of ages under comparison. For the difference of unit between age 25 and 35years, the odds of contracting HIV between age 25 and 35years is exp (10*(0.006179)= 0.03095. This indicates a negative relationship between age and HIV infection and we can say that the odds of contracting HIV decreases by 3% with each additional 10 years increase in age. The inverse relationship between age and Probability of HIV infection suggests that the younger generation below the age of 35years should be of first priority in all the agenda towards eradicating HIV/AIDS spread.

From the table 2 above, the difference in the odds ratio of HIV infection between the male married individuals in the urban population and male single individuals in the urban population is 0.259615. This result implies that the odds the male married individuals in the urban population contracting HIV are 26% higher than that of the male single individuals in the urban population for given age. The difference in the odds ratio of HIV infection between the male married individuals in the rural population and their counterpart is 0.2631593. This result implies that the odds of male married individuals in the rural population contracting HIV are 26% higher than that of the male single individuals for given age.

The odds ratio of HIV infection between female married individuals in urban population is 20% higher than those unmarried female individuals in the same population. Likewise odds of HIV infection between married female individuals in rural population are 21% higher than that of unmarried counterpart.

PREDICTIVE MODEL

We can now predict the likelihood of HIV spread in Kebbi State among different sex, marital status and settlement across all possible age.

1. The predicted model for male married individuals in urban population.

$$\text{Prob.odds MMU} = \frac{\exp(-2.33742 - 0.34753 * Age + 0.24441 + 0.18739 - 0.01356)}{(1 + \exp(-2.33742 - 0.34753 * Age + 0.24441 + 0.18739 - 0.013\,56))} \quad (4.11)$$

2. The predicted model for male married individuals in urban population.

$$\text{Prob.odds MSU} = \frac{\exp(-2.33742 - 0.34753 * Age + 0.24441 - 0.01356)}{(1 + \exp(-2.33742 - 0.34753 * Age + 0.24441 - 0.01356))} \quad (4.12)$$

3. The predicted model for male married individuals in urban population.

$$\text{Prob.odds MMR} = \frac{\exp\left(-2.33742 - 0.34753 * Age + 0.24441 + 0.18739\right)}{(1+\exp(-2.33742 - 0.34753 * Age + 0.24441 + 0.18739))} \quad (4.13)$$

4. The predicted model for male married individuals in urban population.

$$\text{Prob.odds MSR} = \frac{\exp\left(-2.33742 - 0.34753 * Age + 0.24441\right)}{(1+\exp(-2.33742 - 0.34753 * Age + 0.24441))} \quad (4.14)$$

5. The predicted model for female married individuals in urban population.

$$\text{Prob.odds FMU} = \frac{\exp\left(-2.33742 - 0.34753 * Age + 0.18739 - 0.01356\right)}{(1+\exp(-2.33742 - 0.34753 * Age + 0.18739 - 0.01356))} \quad (4.15)$$

6. The predicted model for female single individuals in urban population.

$$\text{Prob.odds FSU} = \frac{\exp\left(-2.33742 - 0.34753 * Age - 0.01356\right)}{(1+\exp(-2.33742 - 0.34753 * Age - 0.01356))} \quad (4.16)$$

7. The predicted model for female single individuals in rural population.

$$\text{Prob.odds FMR} = \frac{\exp\left(-2.33742 - 0.34753 * Age + 0.18739\right)}{(1+\exp(-2.33742 - 0.34753 * Age + 0.18739))} \quad (4.17)$$

8. The predicted model for female single individuals in rural population.

$$\text{Prob.odds FSR} = \frac{\exp\left(-2.33742 - 0.34753 * Age\right)}{(1+\exp(-2.33742 - 0.34753 * Age))} \quad (4.18)$$

SUMMARY, CONCLUSION AND RECOMMENDATIONS

Summary

Bootstrap shows that older people in Kebbistate are more susceptible to HIV infection. This is evidence from the odds of HIV infection among all the socio economic variables under study for a given age. The odds of male married individuals in urban population contracting HIV in Kebbi state are 18% higher than unmarried individuals in the same population. Likewise the odds of male married contracting HIV in the rural population are 17% higher than single male in the population. This is the same with female also, the odds of female married individuals contracting HIV in rural population are 54% higher than that of their unmarried counterpart. Also unmarried females in rural population has 52% lower odds of contracting HIV than that of married female individuals in the same population for a given age. From the jackknife analysis, the odds of contracting HIV of married male and female individuals in urban and rural population in Kebbistate are 26% higher their unmarried counterpart in the same populations for a given age. The odds ratio of HIV infection between female married individuals in urban population is 20% higher than those unmarried female individuals in the same population. Likewise an odd of HIV infection between married female individuals in rural population is 21% higher than that of unmarried counterpart for a given age.

Conclusion

We assess the performance of two re-sampling methods Bootstrap and Jackknife using logistic regression. Both methods estimate the variability of a statistic from the variability of that statistic between subsamples, rather than from parametric assumptions. But it was observed that the bootstrap gives different results when repeated on the same data, whereas the jackknifegives exactly the same result each time. Part of the interest in this research is to have an idea of HIV distribution in the state; therefore the bootstrap is preferred in this type of research. The bootstrap estimate of model prediction bias is more precise than jackknife estimates with linear models.

Relationship and significance of some socio-economic factors and modes of transmission to the HIV/AIDS was studied using HIV data obtained from Federal Medical Centre BirninKebbi. From the analysis obtained from the two re-sampling methods, we can conclude that HIV transmission in Kebbistate is higher among the married couples than single individuals and concentrate more in the rural areas.

Recommendations

The study made the following recommendations.

1. The government should intensify public awareness more on the older age than the youth.
2. Also government should focus more on public awareness in the rural areas.
3. Mother to child transmission should be given attention as married couples are more susceptible to HIV.

REFRENCES

Bobbit A., Steven P., Moonjung C., James A., and Lewrence R., (2007). Application of Bootstrap method in the International Price Program.

Bello A. O. and Oguntolu F. A. (2005) Application of Bootstrap Re-Sampling Method to a Categorical Data of HIV/AIDS Spread across Different Social-Economic Classes. *International Journal of Statistics and Applications 2015, 5(4): 157-168 DOI: 10.5923/j.statistics.20150504.04*

Center for Disease Control and Prevention (CDC), (2015 December) HIV Transmission. www.avert.org/learn-share/hiv-fact-sheets/hiv-transmission

Chong H.Y. (2003). Resampling methods: concepts, applications, and justification. Practical Assessment, Research & Evaluation, 8(19). Available URL http://PAREonline.net/getvn.asp?v=8&n=19.

Demographic and Health Survey. SAHARA-J: *Journal of Social Aspects of HIV/AIDS*, 9:2,5463,DOI:10.1080/17290376.2012.683579.https://doi.org/10.1080/17290376.2012.683579

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7, 1-26.*

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 63, 589-599.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38.

Efron B. & Gong G. (1983). A Leisurely Look at the Bootstrap, the Jacknife, and Cross validation. *The American Statistician*, Vol. 37, No.1 (Feb.,1983), 36-48http://www.sas.rochester.edu/psc/clarke/405/EfronGong.pdf

Efron, B, and Tibshirani, R. J. (1993). *Monographs on, an introduction to the bootstrap.* Statistics and Applied Probability, No. 57. Chapman and Hall, London. 436pp.

Felicia U. I. and Ayodele J.A. (2014), Sociocultural and economic factors influencing the use of HIV/AIDS information by Women in Ugep, Cross River State, Nigeria. *Library Philosophy and Practice (e-journal). 114*

HIV Integrated Biological and Behavioral Surveillance Survey, (IBBSS) 2010. (NACA/FMOH).

John H. (2013). The Jacknife and Bootstrap.URL: http://www.biostat.umn.edu/~johnh/pubh8422/notes/Jackknife_and_Bootstrap.pdf

Joshua K. (2012), Risk factors associated with HIV infection among young persons aged 15–24 years: Evidence from an in-depth analysis of the 2005–06 Zimbabwe. *Journal of Social Aspects of HIV/AIDS, 9:2, 54-63, DOI: 10.1080/17290376.2012.683579*

Kalala N. (2005), The Socio-Economic Factors that fuel Sexually Transmitted Diseases (STDS) and which could result in the Transmission of the Human Immunodeficiency Virus (HIV) in Rustenburg.*Stellenbosch University https://scholar.sun.ac.za*

Krus, D. J. & Fuller, E. A. (1982). Computer-assisted multicross-validation in regression analysis. *Educational and Psychological Measurement*, 42, 187-193.

Manon L. and Andrew J. (2016). A Direct Comparison of Two Densely Sampled HIV Epidemics: *The UK and Switzerland. 1 Scientific RepoRts | 6:32251 | DOI: 10.1038/srep32251 (*www.nature.com/scientificreports)

National HIV/AIDS and Reproductive Health Survey (NARHS Plus II 2012)

Progress report on the Global Plan: *towards the elimination of HIV infection among children by 2015 and keeping their mothers alive*, USAID, 2014).

National HIV Sero-Prevelance sentinel Survey among pregnant women attending antenatal clinics in Nigeria, (FMOH, 2014).

Nigeria HIV/AIDS Indicator and Impact Survey (NAIIS) 2018.

Roposo L.M., Arruda M.B., Brindeiro R.M, Nobre F.F.; Logistic Regression Model for Predicting Resistance to HIV Protease Inhibitor Nelfinerir, XIII Mediterranean Conference on medical and Biological Engineering and Computing 2013,IFMBE proceedings volume 41, 2014,pp 1237-1240. Springer, http://link.spriger.com/chapter/10.1007%2F978-3319-2_306#

Sawyer S. (2005). Re-sampling data: using a Statistical Jacknife. http://www.math.wustl.edu/~sawyer/handouts/Jackknife.pdf

Scholz F.W. (2007) University of Washington, The Bootstrap small sample properties.http://faculty.washington.edu.fscholz/Reports/bootstrap-report.pdf

Scott A. Czepiel, Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation, http://czep.net/contact.html.

Walsh B. (2000) Resampling Methods: Randomization Tests, Jackknife and Bootstrap Estimators.http://nitro.biosci.arizona.edu/courses/EEB581-2006/handouts/random.pdf

Wu C. F (Dec.1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis Source: *The Annals of Statistics*, Vol. 14, No. 4 (Dec., 1986), pp. 1261-1295 URL: http://www.jstor.org/stable/2241454

©2021 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.