

Logistic Regression modelling and Chi Square statistic to predicting chance of survival of STD in Ondo State, Nigeria

J.A. Kupolusi
Department of Statistics, Federal University of Technology, Akure
jakupolusi@futa.edu.ng.

ABSTRACT

Sexually Transmitted diseases are major problems in the health sector. Several researches have shown that it can shorten the lives of people and can cause serious morbidity. This research examined application of Chi Square Statistic and linear logistic regression model to predict the chances of survival among victims of Urinary Tract Infection and Gonorrhoea based on their age, gender and the disease type in Ondo State, Nigeria. Based on the data analysed, we were able to deduce that the male gender irrespective of the age has a greater chance of surviving any sexually transmitted disease. It is therefore recommended that the female sex irrespective of their age undergo constant examination for early detection of any sexually transmitted disease. It is also recommended that there should be sensitization programme for female on sex education irrespective of their age to enable them avoid the infection.

Keywords: Sexually Transmitted Disease, Logistic Regression, Odd Ratio, Urinary Tract Infection and Gonorrhoea.

1.0 Overview On Sexually Transmitted Diseases (STD)

Sexually transmitted diseases are infections that are passed from one person to another through sexual contact. It refers to some set of clinical infections in which the mode of transmission is sexual contact and in which at least one of the partner is infected. Sexually transmitted diseases are recognised as a major public health issue in most industrialised world (centre for disease control and prevention, 2011).

These diseases are caused by bacteria, viruses, yeast and parasites. Among the Sexually transmitted diseases including chlamydia, genital herpes, gonorrhoea, Human immune virus and acquired immune deficiency syndrome, Human papilloma virus, syphilis and trichomoniasis. (centre for disease control and prevention, 2011). Most of these affect both sex but in many cases, the health problem is severe in women (centre for disease control and prevention, 2011). The aim of this research is to use the linear logistic regression model to

predict the chances of survival among victims of Urinary Tract Infection and Gonorrhoea based on their age and gender and the disease type.

1.1 Transmission

The mode of transmission varies among the different sexually transmitted diseases. Some bacteria or virus are found in vaginal secretions or semen (e.g. gonorrhoea), while others are shed from the skin of and around the genitals (e.g. HSV and HPV). Infections typically occur during sexual intercourse or when the genitals come into close contact. Infections may also occur during oral sex. It may also be transmitted during non-consenting sex acts such as rape or molestation. The transmission of a sexually transmitted disease is more efficient from men to women than from women to men. For example, with just one unprotected sexual encounter with an infected partner, a woman is twice as likely to acquire gonorrhoea or Chlamydia. In addition, different sexually transmitted diseases have different rates of transmission. For example, with one unprotected sexual intercourse, a woman has 1% chance of acquiring HIV, 30% chance of acquiring herpes and 50% chance of contracting gonorrhoea if her partner is infected (Mac Donald; Noni, E, David, M. Patrick, 2003). According to (Ayo Adebowale & Musibau T, 2013), Nigeria has a fast growing population and is confronted with numerous health challenges. With a population of 200 million, the country's population is young; therefore, the future of the country rests to a greater extent, on how successful, its youth have a transition to a healthy and productive adulthood.

In this research, two diseases would be considered namely: Urinary Tract Infection (UTI) and Gonorrhoea.

1.1.1Urinary Tract Infection

A urinary tract infection is an infection from microbes, most caused by bacteria but some caused by fungi and in rare cases by viruses. Urinary tract infections are among the most common infections in humans. A Urinary tract infection can happen anywhere in the urinary tract. The urinary tract is made up of kidneys, ureters, bladder and urethra. Most UTIs only involve the urethra and bladder in the lower tract. However, Urinary tract infection can involve the ureters and kidney in the upper tract. Although upper tract UTIs are rare and severe than lower tract UTIs.

1.1.2 Gonorrhoea

Gonorrhoea is a common sexually transmitted disease sometimes referred to as the clap. It affects hundreds of thousands of men and women annually in the country. Globally there is an

estimated 78 million new cases diagnosed each year. However, not all cases are diagnosed and reported; only 333,004 cases of gonorrhoea were reported in the United States in 2013. It is caused by a bacterium called Neisseria gonorrhoeae. Gonorrhoea is easily treated but can cause serious and sometimes permanent complications.

1.2 Application of STD to Logistic regression

Logistic regression is an aspect of statistics that has been widely used to analyse issues on various sectors including machine learning, medical fields, social sciences For example, the Trauma and Injury severity Score (TRISS) which is widely used to predict mortality in injured patients was developed by Boyd et al using logistic regression model (Boyd, Tolson, & Copes, 1987). Many other medical scales used to predict severity of a patient have been developed using logistic regression model, It may also be used to predict the risk of developing a given disease based on observed characteristics of the patient (age, body mass index, sex, test results) (Freedman, 2009). Sexually Transmitted diseases are major problems in the medical sector. For the purpose of this research, logistic regression model will be used to predict the the chances of survival based on age, sex and the type of sexually transmitted diseases contacted in Ondo State.

1.3 Hypothesis Testing

The hypothesis to be tested is given as:-

Hypothesis 1

There is no significant difference between gender and survival rate.

Hypothesis 2

There is no significant relationship between age and survival rate

2.0 Research and Methods

The secondary source of data was employed for the purpose of this research. Data on sexually transmitted diseases with a specific attention on Gonorrhoea and Urinary Tract Infection was collected for the period of January 2015 to December 2017 from the State Specialist Hospital, Akure, Ondo State. The analysis of the data employed Chi Square statistic and logistic regression model. Chi Square statistic was used to test for the relationship between gender and

status of the disease and also age and status of the disease on patients. Logistic regression was used to predict the survival rate of the patients in relation to the disease status, age and gender.

2.1 Chi-Square Test

The chi Square test is calculated by this formula

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Where e_{ij} is the number of cases expected in the ith row of the kth column when the null hypothesis is true and o_{ij} is observed case

Degree of freedom describes the number of values in the final calculation of a statistic that is free to vary and its often denoted by v. The chi square test of significance is given by;

v=(r-1)*(k-1) for h>1 and k<1

where h is the number of rows and k is the number of columns

χ^2=χ^2(r-1)(k-1)

2.2. Logistic Regression Model

The linear logistics regression model was used in the analysis of this paper to determine the factors contributing to the survival of a patient having any of the diseases mentioned.

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analysis, the logistic regression analysis is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one or more nominal, ordinal, interval or ratio- level independent variables. (Dr James Lani, 2018). Logistic regression was developed by David Cox in 1958. The binary logistic is used to estimate the probability of a binary response based one or more independent variables. It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor.

In Statistics, the logistic model is a statistical model that is usually taken to apply to a binary dependent variable. In regression analysis, logistic regression estimates the parameters of a logistic model. More formally, a logistic model is one where the log-odds of the probability of an event is a linear combination of independent variables. The two possible dependent variables values are often labelled as ‘0’ and ‘1’ which represent outcomes such as pass/fail,

win/lose, alive/dead or healthy/sick. The binary logistic regression model can be generalised to more than two levels of the dependent variables.

Categorical inputs with more than two values are modelled by multinomial logistic regression and if the multinomial logistic regression are ordered, by ordinal logistic regression. (Walker & Duncan, 1967).

2.3 Logits and odds Ratio Logits (log)

In statistics, the logit function or log odds is the logarithm of the odds**:** $\frac{\rho}{1-\rho}$ where **ρ** is the probability.

It is also called the standardised logistic regression coefficients. It is a type of probability that creates a map of probability values from **[0,1] to [-∞,+∞]** which is the inverse of the sigmoidal logistic function or logistic transformation.

If ρ is a probability, then $\frac{\rho}{1-\rho}$ is the corresponding odds.

the logit of the probability is the logarithm of the odds:

logit(p) = log($\frac{\rho}{1-\rho}$) = log (p).

The logistic function of any number α is given by the inverse logit:

$logit^{-1}(\alpha)$ **=logistic(α) =**$\frac{e^{\rho}}{1+e^{\rho}}$

The value of the logit is the value of the change in the log odds of the dependent variables per unit per unit change in the predictor variable, positive or negative.

2.4 Odds ratio

The most common way of interpreting a logit is to convert it to odds ratio using the exponential function. It is the ratio of the two odds (The probability of success (response =1) divided by the probability of failure) odds ratio above 1 refer to positive odds although logits can take negative or positive values, odds ratio must be greater than zero. It is a measure of association between exposure and an outcome. The odd ratio represents the odds that an outcome will occur given a particular exposure compared to the odds of the outcome occurring in the absence of the exposure. The Facts about Odds ratio include:-

1. It is calculated in case control studies as incidence of outcome is not known.
2. Odds ratio >1 indicates increased occurrence of events.
3. Odds ratio <1 indicates decreased occurrence of event.

Mathematically, the odd ratio is given as Odds =$\frac{\rho}{1-\rho}$ (1)

This is the ratio of the probability of 1 to probability of 0. Unlike probability, the odds ratio can be greater than 1. The natural log of the odds ratio is the logit ρ. Thus

Logit(ρ)=ln$(odds)$=ln$(\frac{\rho}{1-\rho})$ (2)

Logit (ρ) ln$(odds)$ = ln$(\frac{\rho}{1-\rho})$ = $\beta_0 + \beta_1 x_1$ (3)

Taking exponential, we obtain

$\theta(x)$=$(\frac{\rho(x)}{1-\rho(x)})$=exp($\beta_0 + \beta_1 x_1$) (4)

Where exp =e=2.718 is the base of the natural logarithm.

Solving for $\theta(z)$, we obtain

P(x)=$\frac{\exp(\beta_0+\beta_1 x_1)}{(1+\exp(\beta_0+\beta_1 x_1)}$ (5)

From equation (5), the probability of response is given by P(Y=1) = $\mu = \frac{e^{\eta(x)}}{1+e^{\eta(x)}}$ (6)

where

$$\eta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

β_i is the parameter to be estimated (i =1,2,3)

x_1 = Age (male =1 and Female = 0)

x_2 = Sex (young = 1 and old = 0)

x_3 = Disease type (Urinary Tract Infection =1 and gonorrhoea = 0)

$\eta(x)$ = (Treated = 1 and complicated = 0)

3.0 Data Analysis and Result

3.1 Chi-Square

Hypothesis:

H_0: there is no relationship between gender and Status

H_1: there is a relationship between gender and Status

Table 3.1: Cross tabulation on the association between sex and status of patient

Status		Sex		Total
		Female	Male	
Complicated	Count	456	181	637
	Expected count	374.3	262.7	637
	Residual	81.7	-81.7	
Treated	Count	714	640	1354
	Expected	795.7	558.3	1354
	Residual	-81.7	81.7	
	Count	1170	821	1991
	Expected count	1170.0	821.0	1991.0

	Value	Df	Asymp. sig. (2-sided)	Exact sig.(2-sided)	Exact sig. (1-sided)
Pearson Chi-square	63.541[a]	1	.043	.062	.038
Continuity Correction[b]	62.766	1	.072		
Likelihood Ratio	65.268	1	.054	.062	.038
Fisher's Exact Test				.062	.038
Linear- by- linear Association	63.510	1	.053	.062	.038
N of Valid Cases	1991				

Table 3.1 above shows the Chi-Square Test on the relationship between sex and status of patient using the Pearson Chi-square probability value P-value=0.043

Decision Rule: Reject H_0 if P-value<α, else do not reject

Conclusion: Since the P-value < 0.05 that is 0.043 < 0.05 at 5% level of significance, we reject H_0 and conclude that sex is a determinant factor of surviving gonorrhoea and Urinary Tract infection.

Table 3.2: Cross tabulation on the association between age and status of patient

Status		Age		Total
		Old	Young	
Complicated	Count	526	111	637
	Expected count	459.8	177.2	637.0
	Residual	66.2	-66.2	
Treated	Count	911	443	1354
	Expected	977.2	376.8	1354.0
	Residual	-66.2	66.2	
	Count	1437	554	1991
	Expected count	1437.0	554.0	1991.0

	Value	Df	Asymp. sig. (2-sided)	Exact sig.(2-sided)	Exact sig. (1-sided)
Pearson Chi-square	50.445[a]	1	.054	.016	
Continuity Correction[b]	49.686	1	.072	.016	
Likelihood Ratio	53.328	1	.052	.018	
Fisher's Exact Test				.018	
Linear-by-linear Association	50.419	1	.054	.003	.003
N of Valid Cases[b]	1991				

Table 3.2 shows the Chi-Square Test on the relationship between age and status of patient

Using the pearson Chi-square probability value P-value=0.054

Decision Rule: Reject H_0 if P-value<α, else do not reject

Conclusion: Since the P-value > 0.05 that is 0.054 > 0.05, we do not reject H_0 and conclude that age is not a determinant factor of surviving gonorrhoea and Urinary Tract infection.

3.2 Logistic Regression analysis

Table 3.3**:** Dependent variable encoding

Original Value	Internal Value
Complicated	0
Treated	1

Table 3.4: Variable in the equation

		B	S.E	Wald	df	Sig	Exp(B)
Step B	Constant	.754	.048	246.312	1	.000	2.126

Table 3.5 Classification table[a]

	Observed		Predicted		
			Status		Percentage correct
			Complicated	Treated	
Step 1	Status	Complicated	41	596	6.4
		Treated	48	1306	96.5
	Overall percentage				67.7

a. The cut value is .500

Table 3.6: Variables in the Equation

		B	S.E	Wald	df	sig	Exp(B)	95.0% C.I for Exp(B)	
								Lower	Upper
Step 1[a]	Age(1)	-.945	.122	59.681	1	.658	2.573	2.024	3.270
	Sex(1)	.905	.106	73.345	1	.001	2.473	2.010	3.024
	Disease(1)	-.302	.145	4.334	1	.054	.740	.557	.982
	Constant	.222	.070	9.977	1	.002	1.248		

Logistic Regression of the status(Y) on the independent variable Sex (X_1), Age (X_2) and Disease (X_3).

From the analysis above, it was shown that the age and type of disease have negative coefficient which implies that that they have little or no effect on the survival of the patient while sex has a positive coefficient which means that it is contributing to the survival of patients in ondo state.

The predictive equation is

$$P(Y=1) = \mu = \frac{e^{(0.222-0.945x_1+0.905x_2-0.302x_3)}}{1+e^{(0.222-0.945x_1+0.905x_2-0.302x_3)}}$$

Odds Ratio $= e^B$

The odds ratio is greater than one in sex category that is; 2.4719>1 which tells us that the male is more likely to survive than the female

4.0 Conclusion

The logistic regression model shows that sex is significant in predicting the status of the patient after visiting the hospital while age and diseases are not significant in predicting the status therefore the male has a higher chance of survival than the female irrespective of the age and patient infected with either gonorrhoea or Urinary Tract Infection have equal chance of survival as disease type is insignificant to outcome of survival.

Based on the data analysed, it was deduced that the male sex has a higher chance of surviving after visiting the hospital and lower chance of even getting infected with the disease. The age and type of disease a patient is infected with does not determine the outcome of the survival of the patient.

References

Adebowale,A. S., & Musibau, T. A. F. (2013). *Statistical modelling for social risk factor for sexually transmitted diseases among female youths in Nigeria.* badan: Department of epidemology.

Baltas, G.,& Doyle, P. (2001). Random Utility Models in Marketing Research. *A Survey*.

Belsley David. (1991). Collinearity and weak data in regression. *Conditioning and Diagnostics*.

Boyd, C. R., Tolson, M. & Copes, W. S. (1987). Evaluating trauma care: The TRISS method. *The Journal of Trauma*.

Freedman, D. A. (2009). *Statistical Models: Theor and practice* . Cambridge University Press.

Linda, A.C. (2011). *Modelling of sti prevalence hiv infected adults in hiv care programs in kenya using logistic regression.* university of nairobi. (n.d.). *log odds ratio.* nist.gov.

Mac, D., Noni, E., David, M. & Patrick. (2003). Seexually transmitted disease syndromes. *Principles and practice of infectious diseases*.

Marissa R. Gray. (2014). Risk Factor for sexually transmitted Bacterial and Viral diseases. *analysis of 2007-2010 nhanes data*, 48.

Oates, K. S. (2001). *logistic regression on score sending and college matching among high school students.*

Prevention, C. f. (2011). *Sexually transmitted disease surveillance.* Atlanta: U.S Department of Health and Human Services.

Prevention, Centre For Disease and Control and. (2011). *Sexually transmitted disease surveillance.* Atlanta: U.S Department of Health and Human Services.

Saari, Donald G. (2001). Decisions and Elections. *Explaining the unexpected.*

U.S National Library of Medicine. (2018). *Sexually transmitted diseases.* Bethseda: National Institute of Health.

Walker, S., & Duncan, D. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*.

