International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-8, Issue-7, July 2022
ISSN: 2395-3470
www.ijseas.com

Using Regression Coefficient in Epidemiology; A Case Study of HIV

Ejiro.S.Omokoh and Ifeoma.O.Ejinkonye

Department of Mathematics and Computer Science, Western Delta University, Oghara, Delta State, Nigeria.

Abstract

However, the majority of these models have been implemented with assumptions that may be limiting or completely lead to less meaningful outcomes and therefore interpretations. These assumptions include the linearity, stationarity, and normalcy assumptions. According to research, the linearity assumption does not hold true for all factors. Age, for example, has been found to have a non-linear connection with HIV prevalence. Other studies have assumed the stationarity assumption, which states that a single stimulus, such as education, elicits the identical response in all of the regions under consideration, which is likewise highly limited. Responses to stimuli may differ from location to region due to factors such as culture, preferences, and attitudes.The stationarity assumption is relaxed by using the conditional coefficients of regression derived from the autoregressive model to allow the rest of the covariates to vary spatially, and the linearity assumption is relaxed by using the random walk model of order 2 to allow the covariate age to have a non-linear effect on HIV prevalence.

1.0 Introduction

The first incidence of AIDS in Nigeria was discovered in 1985 in a young female teenager aged 13 years, but it was not publicized until 1986. This example was discovered in Lagos, Nigeria's former capital and largest metropolis (Nasidi et al, 1980). The revelation of the presence of AIDS in Nigeria was met with skepticism and scepticism by the Nigerian people. AIDS was viewed as a disease of American homosexuals, an illness from a faraway land with no place in Nigerian society. People, particularly young people, were suspicious of the presence of HIV in their surroundings. They saw the whole story as a hoax and a ploy by the Americans to discourage sex(Oyefara, 2007). This mistrust was ingrained in the multiple AIDS acronyms, one of which was "American Idea for Discouragement of Sex." The administration was then more concerned in debating the disease's origins, even denying that the disease posed a threat to the Nigerian country. The government and the general population took this approach since the first HIV-positive person diagnosed in the country was a sex worker from one of the West African countries.

Because of the public's understanding of the disease, as well as most Nigerians' religious and cultural beliefs that death is pre-ordained and must occur when it is due, there has been little or no behavioral change in areas of sex and sexual practices (Adeokun, 2006). As a result, the AIDS virus moved silently and unnoticedly throughout the country's sexual networks, affecting all socioeconomic classes, professions, age groups, genders, regions, zones, states, cities, and villages (Orubuloye and Oguntimehin , 2006).When the HIV/AIDS epidemic in Nigeria was initially assumed to be limited to a few subpopulations, it quickly developed into a widespread pandemic (UNAIDS , 2006). Twenty-two years after the first case was reported, the disease has

spread to become a large epidemic that is not just a health burden but also a socioeconomic issue. It has impacted every aspect of Nigerian society and has eaten its way deep into the Nigerian country.

Extramarital and nonmarital sexual encounters are one of the major causes behind this pandemic in Nigeria. According to the 2003 National HIV/AIDS and Reproductive Health Survey (NARHS), around 9% of women aged 15-49 and 18.4% of males aged 15-64 engage in extramarital and premarital sexual activity. The survey also indicated that youngsters are more vulnerable, with approximately 14% of female and 25% of male youth engaging in non-marital sex. Furthermore, only about 32% of women and 50% of men use condoms during unsafe sex.

In the past, most HIV statistics were based on assumptions. This is dangerous since those assumptions rarely produce acceptable results and frequently miss the mark because certain data points cannot be gathered owing to the size of the datasets. The regression coefficient is used in this study as a tool for making predictions about HIV occurrence and spread in Nigeria in order to determine the relationship between various entities such as age, gender, geography, and so on and how they influence the rate of HIV spread in Nigeria.

2.0 Regression coefficients

Regression coefficients describe the relationship between a predictor variable and the response by providing estimates of unknown population parameters. Coefficients are the values that multiply the predictor values in linear regression (Jim Frost, 2014). Consider the following regression equation: $3x + 5 = y$. In this equation, +3 represents the coefficient, x represents the predictor, and +5 represents the constant. The direction of the link between a predictor variable and a responder variable is shown by the sign of each coefficient. A positive sign suggests that when the predictor variable increases, so does the responder variable. A negative sign suggests that the response variable drops as the predictor variable increases.

The coefficient value represents the mean change in the response when the predictor is changed by one unit. For instance, if a coefficient is +3, the mean response value rises by 3 for every unit change in the predictor.

Binary variables are widely used in statistics to estimate the likelihood of a given class or event occurring, such as the likelihood of a team winning, a patient being healthy, and so on, and the logistic model has been the most regularly used model for binary regression since around 1970 (Cramer, 2002). When there are more than two possible values for binary variables (e.g., whether an image is of a cat, dog, lion, etc.), binary logistic regression can be generalized to multinomial logistic regression. When the various categories are ordered, ordinal logistic regression (for example, the proportional odds ordinal logistic model (Walker, 1967)) can be used.The logistic regression model simply models the probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to create one, for example, by selecting a cutoff value and classifying inputs with probability greater than the cutoff as one class and those with probability less than the cutoff as the other; this is a common way to create a binary classifier.

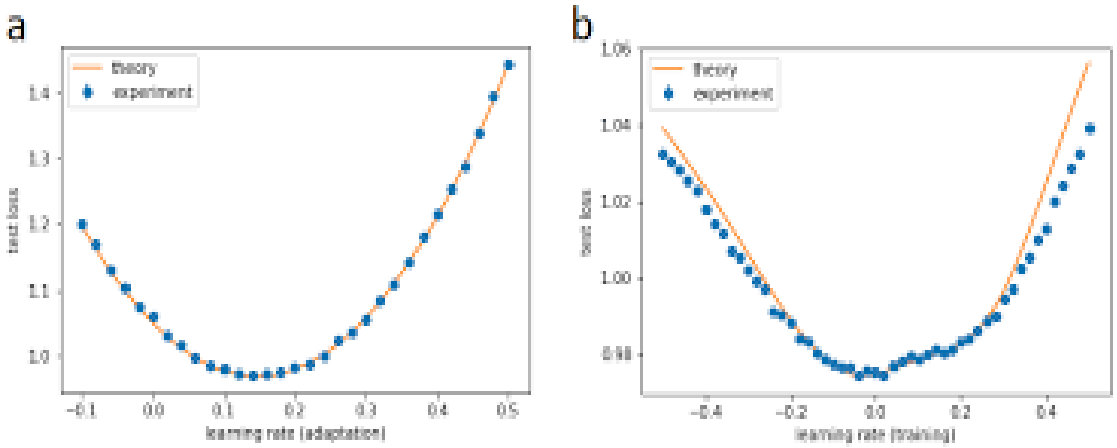

Figure 2.1 shows regression curves

2.1 Machine and Deep Learning

Machine learning is a branch of artificial intelligence that tries to use intelligent software to enable computers to do skilled tasks. The backbone of intelligent software that is utilized to generate machine intelligence is statistical learning methods. Because machine learning algorithms require data in order to learn, the field must be linked to database science. Similarly, words like Knowledge Discovery from Data (KDD), data mining, and pattern recognition are common. One might ponder how to view the big picture in which such a link is shown.

Anthony C et al,(2020), both bioinformaticians, submitted a commentary in -Based Medicine, 2020, offering an intelligent Intelligence strategy to make DL more interpretable by focusing on feature scoring and data synthesis (DS). Logistic regression is faster than other supervised classification approaches such as kernel SVM or ensemble algorithms, but it is less accurate. It suffers from the same limitations as linear regression in that both techniques are simply too basic for complex correlations between variables. Finally, when the decision boundary is nonlinear, logistic regression tends to underperform. Anthony C et al,(2020), both bioinformaticians, wrote a commentary in -Based Medicine, 2020, offering an intelligent Intelligence strategy to make DL more interpretable by focusing on feature scoring and data synthesis (DS). Logistic regression is faster than other supervised classification approaches like kernel SVM or ensemble algorithms, but it is less accurate. It also has the same issues as linear regression in that both techniques are significantly too basic for complex correlations between variables. Finally, logistic regression tends to underperform when the decision boundary is nonlinear.

2.2 SVMs versus logistic regression

SVMs, like logistic regression, can be generalized to categorical output variables with more than two values. However, the kernel method can also be used for logistic regression (this is known as "kernel logistic regression"). While logistic regression, like linear regression, uses all data points, points near the margin have significantly less influence due to the logit transform, and so, despite the math being different, they frequently produce results comparable to SVMs (Prakash Nadkarni, 2016).

When deciding between SVMs and logistic regression, it is generally advantageous to use both. SVMs sometimes provide a better fit and are more computationally efficient than logistic regression—logistic regression utilizes all data points but subsequently discounts the values outside the margin, whereas SVM uses only the support-vector data points to begin with. However, in terms of interpretability, SVM is a bit of a "black box." However, with logistic regression, the contribution of individual variables to the final fit can be better understood, and the outputs of data back-fitting can be directly evaluated as probabilities.

2.3 Multinomial Logistic Regression

Multinomial logistic regression (MLR) is a semiparametric classification statistic that extends logistic regression to issues with multiple classes (e.g., more than two possible outcomes). MLR forecasts the probabilities of possible outcomes of a categorically distributed dependent variable using any class of independent factors (e.g., binary, ordinal, continuous). MLR predicts group membership using the log odds ratio rather than probabilities, and the final model is fitted using an iterative maximum likelihood method rather than a least squares method. MLR makes the following assumptions: (1) each independent variable should have a single value for each case; (2) collinearity is considered to be reasonably low, albeit not necessarily totally independent; and (3) irrelevant alternatives should be independent (IIA). IIA defines the odds of preferring one class over another, regardless of the presence of other unrelated and irrelevant alternatives.Unfortunately, MLR does not generate typicality probabilities that can be used to assess how well the model classifies, although Hefner and Ousley (2014) propose replacing these metrics with nonparametric methods such as ranked probabilities and ranked interindividual similarity measures (Hefner and Ousley 2014). Cross-validation methods can be used to test the performance of logistic regression and determine the model's dependability.

3. 0 Methodology

The data for this study came from the Nigerian AIDS Indicator Survey (KAIS), which was conducted by the Nigerian government with funding from the US President's Emergency Plan for AIDS Relief (PEPFAR) and the United Nations (UN). The primary goal of the survey was to collect high-quality data on the prevalence of HIV and Sexually Transmitted Infections (STI) among adults, as well as to assess the population's understanding about HIV and STIs.

The National Sample Survey and Evaluation Programme IV served as the sample frame for KAIS (NASSEP IV). It was made up of 1800 clusters, 1260 rural and 540 urban, with 294 rural and 141 urban clusters sampled for KAIS. KAIS 2007's overall design was a stratified, two-stage cluster sampling design. The first stage entailed picking clusters from NASSEP IV, and the second stage involved selecting households for KAIS with equal probability in the districts'

urban-rural strata. For KAIS, a sample of 415 clusters and 10,375 households were methodically chosen. An equal probability systematic sampling method was used to select a uniform sample of 25 households per cluster.

The survey had two parts: A household questionnaire was used to collect information on the living environment, and an individual questionnaire was utilized to obtain information about demographic features and HIV and STI awareness on men and women aged 15–64 years. A representative sample of families and individuals was drawn from the country's eight provinces. Each person was requested to produce a venous blood sample for HIV testing. The final KAIS, 2007 report contains more information on survey methodology utilized in data collection (NASCOP, 2007).Despite the fact that a new round of KAIS, 2012 (NASCOP, 2012) has been completed, this study uses data from 2007. Because the final release of this new data had not been done, the data was not usable. The women's data from the KAIS 2007 survey were used in this study. The analysis employed data from 4864 women aged 15–64 years who had donated venous blood for HIV testing and also had full covariate information.

3.1 Statistical model for the HIV

The relevance of the covariates was determined by fitting a univariate standard logistic model between each single covariate and the outcome variables (HIV status). At the 5% threshold of significance, the connection was deemed significant. Table 1

Let y_{ijk} be the disease k status (0/1), $k = 1$ for HIV and $k = 2$ for HSV-2, for individual j in county i: $i = 1, 2, \ldots, 46$. $y_{ij1} = 1$ if individual j in county i is HIV positive and zero otherwise and $y_{ij2} = 1$ if individual j in county i is HSV-2 positive and zero otherwise. This study assumes the dependent variable y_{ij1} and y_{ij2} are univariate Bernoulli distributed, i.e. $y_{ij1}|p_{ij1} \sim \text{Bernoulli}(p_{ij1})$ and $y_{ij2}|p_{ij2} \sim \text{Bernoulli}(p_{ij2})$.

The p continuous independent variables are contained in the vector $\mathbf{X_{ijk}} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})'$ while $\mathbf{W_{ijk}} = (w_{ij1}, w_{ij2}, \ldots, w_{ijr})'$ contains r categorical independent random variables with the first component accounting for intercept. In this study, $p = 1$ (age) and $r = 8$.

The unknown mean response namely $E(y_{ijk}) = p_{ijk}$ (Manda, 2007)relates to the independent variable as follows:

$$h\left(p_{ij1}\right) = \boldsymbol{X^T\beta_1} + \boldsymbol{W^T\gamma_1}, \text{ for HIV and}$$

Where h(.) is a logit link function, is a p-dimensional vector of regression coefficients for continuous independent variables, and is an r-dimensional vector of regression coefficients for categorical independent variables. To account for the non-linear effects of the continuous covariates as well as the spatial autocorrelation in the data, a random walk model of rank 2 (RW2) and a convolution model were used.

The RW2 model approach relaxed the highly restrictive linear predictor by a more flexible semi-parametric predictor, defined as:

$$h\left(p_{ij1}\right) = \sum_{t=1}^{p} f_t\left(x_{ijt}\right) + f_{spat}\left(s_{i1}\right) + \boldsymbol{W^T\gamma_1} \text{ for HIV}$$

The function $f_t(.)$ is a non-linear twice differentiable smooth function for the continuous covariate and $f_{spat}(s_{ik})$ is a factor that caters for the spatial effects of each county. This study utilized the convoluted spatial structure which assumes that the spatial effect can be decomposed into two components: spatially structured and spatially unstructured i.e. $f_{spat}(s_{ik}) = f_{str}(s_{ik}) + f_{unstr}(s_{ik})$, $k = 1, 2$ (Ngesa et al, 2014).

4.0 RESULT

According to the data, the proportion of girls who tested positive for the virus is

higher than that of males. Females account for around 61 percent of all positive cases, with males accounting for the remaining 39 percent. Females were also infected at an earlier age than males. The majority of infections were discovered in people aged 15 to 49 years. This age group accounted for approximately 86 percent of all female cases and approximately 78 percent of male cases. Females aged 25-34 years were the most impacted by the infection. Those aged 35 to 49 were slightly more affected than males aged 25 to 34.This appears to indicate that older men infect younger females. According to a formal study of the data using the logistic regression model, there is a three-way interaction between time, age, and gender. A simpler, less complex model was sought, but it appears that the three-way interaction model is the best. The data was then partitioned by time to exclude the effect of time and evaluate the two-way interaction of sex and age for each time period. This analysis indicated that the two-way interaction was significant in both time periods1. As a result, the likelihood of testing positive for HIV infection is determined by the patient's gender and age.

Variable	*P*-Value	Unadjusted OR
Demographic characteristics		
Place of residence (Ref Rural)		
Urban	0.001	0.749 (0.635, 0.884)
Age (Ref 15–19)	0.000	
20–24	0.000	2.825 (1.982, 4.026)
25–29	0.000	3.055 (2.133, 4.375)
30–34	0.000	4.656 (3.276, 6.618)
35–39	0.000	3.682 (2.544, 5.328)
40–44	0.000	2.796 (1.869, 4.181)
45–49	0.000	2.783 (1.858, 4.169)
50–54	0.000	2.347 (1.490, 3.696)
55–59	0.294	1.352 (0.770, 2.375)
60–64	0.173	0.487 (0.173, 1.371)
Social Characteristics		
Wealth Quantile (ref poorest)	0.525	
Second	0.652	1.058 (0.827, 1.353)
Middle	0.392	0.896 (0.696, 1.153)
Fourth	0.564	1.074 (0.843, 1.369)
Richest	0.592	0.938 (0.741, 1.186)

Table 1 Exploratory data analysis for HIV for women in Nigeria

4.1 Further Discussion

Although the SVC model for HIV did not change significantly from its stationary version, the choropleth maps indicate that the impact of some of the factors varied over space. Education had a greater impact on HIV prevalence among women in the North Eastern, Coastal, Southern, and Central regions, as indicated. Age at first sex also had a bigger influence in areas where education had a greater impact than in other parts of the country, indicating a link between education and age at first sex. Except for some portions of Nigeria's West and Central regions, where the influence was stronger, the effect of the number of partners had in the previous year was nearly the same across the country.The effect of frequency of away trip was also noticeable in the Eastern, Coastal, and Southern regions, as well as parts of the Central region, but married status was prominent in the Western region.

5.0 CONCLUSION

The goal of this thesis was to create epidemic models that could describe and predict the Nigerian HIV/AIDS epidemic. To do this, we concentrated on two major approaches: spatial epidemiology and back-calculation methods. The nature of the provided data influenced the selection of these strategies. Following a thorough examination of all available HIV/AIDS data sources, two sets of data obtained from two separate sources were determined to be superior to others based on the criteria of nationwide coverage and lowest reporting delay.The data used in this study were the results of a survey of 1057 health and laboratory facilities (public and private) conducted by the Nigerian Institute of Medical Research (NIMR) in 2000, as well as the results of the Federal Ministry of Health's biannual National HIV/AIDS Sentinel Surveillance Survey conducted between 1991 and 2005.

The models presented in this paper allow for the relaxation of two limiting assumptions in disease mapping. The effects of variables on HIV were found to be geographically variable. For example, the effect of education on HIV status was lower in North Eastern and sections of the region than in most other parts of the country. Because age was discovered to have a non-linear effect on HIV prevalence, a linearity assumption would have resulted in incorrect data and interpretations. The findings are important because they can be used to create tailored HIV prevention efforts in various countries. The methods utilized in this study could be repeated in other studies using similar data.

REFERENCES

Adeokun L.(2006): Social and Cultural factors affecting the HIV epidemic: In AIDS in Nigeria: A nation in the Threshold. Harvard Centre for population and Development Studies.

Cramer, J. S. (2002). The origins of logistic regression (PDF) (Technical report). Vol. 119. Tinbergen Institute. pp. 167–178.

Joseph T. Hefner, Kandus C. Linde,(2018): in Atlas of Human Cranial Macromorphoscopic Traits.

Manda O, Leyland H(2007). An empirical comparison of maximum likelihood and Bayesian estimation methods for multivariate disease mapping. S Afr Stat J. pp1–21.

NASCOP (2012): Ministry of Health, Kenya: Kenya AIDS Indicator Survey report.

NASCOP (2007): Ministry of Health, Kenya: Kenya AIDS Indicator Survey report. 2007.

Nasidi A, Harry T O, Ajose-Coker O O, (1986). Evidence of LAV/HTLV III infection and AIDS related complex in Lagos. Nigeria. II international Conference on AIDS, Paris France, June 23-25, FR86–3.

Ngesa O, Mwambi H, Achia T. (2014): Bayesian spatial semi-parametric modeling of HIV variation in Kenya. PLoS One. pp 50-62.

Oyefara J. L. (2007): Food insecurity, HIV/AIDS pandemic and sexual behaviour of female commercial sex workers in Lagos metropolis, Nigeria. Social Science Med. Social Aspects of HIV/AIDS, 4:626–635.

Orubuloye IO and Oguntimehin F. (1999): Death is pre-ordained, it will come when it is due: attitude of men to death in the presence of AIDS in Nigeria. Resistances to Behavioural change to Reduce HIV/AIDS infection, pages 101–111.

Prakash Nadkarni (2016), Core Technologies: Machine Learning and Natural Language Processing in Clinical Research Computing, pp 35-40

UNAIDS(2004): Epidemiological fact sheets on HIV/AIDS and Sexually Transmitted Infections: Nigeria. www.unaids.org/hmtl/pub/publications/factsheets01/Nigeria,.

Walker, SH; Duncan, DB (1967). "Estimation of the probability of an event as a function of several independent variables". Biometrika. **54** (1/2): 167–178.