Simultaneous modeling of Binary Responses: A sequence of binary models and Ordinal multinomial with Bayesian estimates impacting HIV/AIDS

Adriana Dornelles
Arizona State University
Di Fang
University of Arkansas
Ying Wang
Banner Good Samaritan Medical Center
Jeffrey Wilson (jeffrey.wilson@asu.edu)
Arizona State University https://orcid.org/0000-0001-9072-1714

Research article

Keywords: multivariate responses, simultaneous responses, joint modeling

Posted Date: April 25th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-19135/v1

SIMULTANEOUS MODELING OF BINARY RESPONSES: A SEQUENCE OF BINARY MODELS AND ORDINAL MULTINOMIAL WITH BAYESIAN ESTIMATES IMPACTING HIV/AIDS

ADRIANA DORNELLES

DEPARTMENT OF ECONOMICS

ARIZONA STATE UNIVERSITY TEMPE AZ 85287-9801

Adriana.Dornelles@asu.edu

DI FANG

Agribusiness

University of Arkansas Fayetteville Arkansas

difang@uark.edu

YING WANG

Banner Health

wangying9@gmail.com

and

JEFFREY R. WILSON

CORRESPONDING AUTHOR

Department of Economics

Arizona State University

jeffreyw@asu.edu

(480) 213 4460 http://www.springer.com/us/book/9783319238043

http://www.public.asu.edu/~jeffreyw

SIMULTANEOUS MODELING OF BINARY RESPONSES: A SEQUENCE OF BINARY MODELS AND ORDINAL MULTINOMIAL WITH BAYESIAN ESTIMATES IMPACTING HIV/AIDS

ABSTRACT

BACKGROUND: In this research, we examined several binary factors impact binary outcomes simultaneous and how the information of HIV/AIDS is perceived by the public is associated with outcomes to HIV/AIDS.

METHODS: We used polytomous responses through a sequence of binary models and a multinomial logistic regression model with Bayesian estimates to analyze the 2009 Mozambique survey data as it pertains to *blood test*, *heard of HIV/AIDS* and *heard about campaign*.

RESULTS: The analysis reveals that both heard about HIV and heard about the campaign are represented differentially in testing positive. Wealth, education and thinking of risk is positively associated with heard about HIV and heard about the campaign regardless of HIV. However, religious is a positive factor for social efforts of hearing of HIV/AIDS and the campaign. Both the polytomous response model and the ordinal model with model gave the same findings in regards to the marginal mean. However, the polytomous (conditional) models gave additional information about education.

CONCLUSIONS: While knowledge of the disease continues to be important, the future social effort to combat HIV in Mozambique may need different strategies in different subpopulation groups.

Keywords: *multivariate responses; simultaneous responses; joint modeling*

BACKGROUND

In the analysis of survey data, the classical approach is to fit a logistic regression model and to use maximum likelihood, where the inferences and interpretations about the regression coefficients are usually based on the asymptotic theory. The accuracy of the classical methodology is questionable for data in which the subpopulations have random zeros or small frequencies. In such situations, Bayesian methods may provide more accurate estimates due to the advantage of the data augmentation through simulations, Efron (2015). Although many articles have compared results using classical and Bayesian approaches, few studies have made use of the Bayesian model in the analysis of the HIV/AIDS survey data as a depleted and unique set of data.

In the absence of relevant prior experience, popular Bayesian estimation techniques usually begin with some form of 'uninformative' prior distribution intended to have minimal inferential influence. The Bayes rule still produces reliable estimates and credible intervals, but these lack the logical force that is attached to experience-based priors and require further justification, Efron (2015).

In this paper, we present two distinct models to the data with certain unusual subpopulations: a sequence of binary models for polytomous responses and a Bayesian ordinal logistic regression using the Mozambique AIDS/HIV survey data with random zeros. This paper is organized as follows. Section 2 outlines the Mozambique Survey data. Section 3 describes the statistical methods. Section 4 provides the results, and Section 5 gives a summary of key findings and conclusions.

METHODS

OVERVIEW OF THE SAMPLE

Mozambique is a southeast African country that has about 21 million people. In 2008, 11% of population (about 1,600,000) live with HIV (Human Immunodeficiency Virus)

and 56% of them are women and children [Audet, Burlison, Moon, Sidat, Vergara, and Vermund 2010]. The main transmission path for HIV in Mozambique is heterosexual intercourse [Risk analysis. Anal Afr. 1996;], but there are other factors involved in spreading the virus such as unclean needle injection, and blood transfusion. To reduce the prevalence of HIV within Mozambique, numerous world organizations and the Mozambique government have put much effort to combat the disease. These efforts include educating the population regarding the HIV infection and prevention. Additional factors such population demographics, social economic status, and religion are also important to manage HIV/AIDS epidemic in Mozambique [Font, Puigpinos, Chichango, Cabrero, Borrell 2006]. While it is well documented that these campaigns are efficient ways to help change sexual attitude, practices, and risk behavior that consequently prevent and reduce HIV transmission, less is known on how perceiving social efforts are associated with other explanatory factors of HIV.

The Mozambique survey 2009 data provide information about HIV infection, behavioral and social risk factors. One of the main responses was HIV infection, which was measured by blood test. The survey also includes socio-demographic information such as age, sex, education level, work status and marital status. Socioeconomic variables included whether the participant has access to safe drinking water, electricity and has a refrigerator at home. The survey also collected information whether a participant is aware of the educational campaigns about the HIV/AIDS prevention, and finally, whether a participant has knowledge of HIV/AIDS. Thus, the main responses of interest in this paper are measured as: *blood test (responsibility)*, a binary response with value "1" if tested positive for HIV+ and "0" otherwise; *awareness of HIV:* binary variable with value "1" indicating if a person is aware of an educational campaign to prevent HIV and "0" otherwise; *knowledge of HIV:* a binary variable with value "1" indicating

that a person has some knowledge of HIV/AIDS through different sources (e.g. community meetings) and "0" otherwise.

For this study, variables such as marital status, education, work, and electricity in the home are key factors, and therefore were used as covariates in the models and coded as follows: gender (0 females, 1 males), education (0 for ≤ 3 years of education, 1 otherwise), marital status (0 living alone, 1 otherwise), working (0 for not working, 1 otherwise) and electricity at home (0 for no, 1 for yes).

STATISTICAL METHODS

It is customary in any national survey to have a plethora of factors impacting several outcomes of interest in an attempt to identify the key factors. However, those factors may provide different results if they examine the influence on one outcome at a time versus several outcomes. As such, it is important to use models that address the questions based on simultaneous modeling. To do otherwise, may result in make conclusions and recommendations that may not be appropriate. It denies one the opportunity to make statements about the entire system of variables. Due to the uniqueness of the data, we relied on two different simultaneous modeling to identify the key factors.

POLYTOMOUS RESPONSES BY SEQUENCE OF BINARY MODELS

The uniqueness of the Mozambique survey data affords a method of modeling polytomous responses through a sequence of binary models or as a nested dichotomy model. The approach is attractive when the response is naturally arranged as polytomous responses by a system of nested dichotomous models. This method decomposes a multi-class problem into a collection of binary problems. Such a system recursively applies binary splits to divide the set of classes into two subsets and trains a binary classifier for each split, Leathart, Pfahringer, and Frank, (2016). Therefore, the

main predictor of exposure (knowledge) is factored into a sequence of binary choices (trees) and then model them at each node or stage, Figure 1.

FIGURE 1 ABOUT HERE

Based on the data structure shown in Figure 1, we performed three submodels: *Model # 1* is fitted at stage 1 to model the logit of knowledgeable of HIV/AIDS defined as:

$$\textit{Model \# 1}: \text{logit}\,(P_{Know}) = \log\left(\frac{P_{Know}}{P_{\overline{Know}}}\right) = \omega_1 + \beta_1 X_{Gen} + \beta_2 X_{Ele} + \beta_3 X_{Edu} + \beta_4 X_{Emp} + \beta_5 X_{Mar},$$

where β_i for $i = 1,2,\ldots,5$; X_{Gen} denotes the covariate for gender, X_{Ele} denotes the covariate for electricity, X_{Edu} denotes the covariate for education, X_{Emp} denotes the covariate for employment, and X_{Mar} denotes the covariate for marriage; P_{Know} denoted the probability of knowledgeable about HIV/AIDS and $P_{\overline{Know}}$ is the complement. *Model #2* is fitted at stage 2, to model using data from those subjects that are knowledge about HIV/AIDS, to model the logit of being aware of the HIV/AIDS campaign:

$$\textit{Model 2}: \quad \text{logit}(P_{Aware}) = \omega_1 + \beta_1 X_{Gen} + \beta_2 X_{Ele} + \beta_3 X_{Edu} + \beta_4 X_{Emp} + \beta_5 X_{Mar}$$

where P_{Aware} denoted the probability of being aware of the campaign about HIV/AIDS and $P_{\overline{Aware}}$ is the complement. The *Model #3* is fitted at stage 3, to model using data from those subjects that are knowledge about HIV/AIDS and are aware of the HIV/AIDS campaign, to model the logit of a positive blood test for HIV/AIDS.

$$\textit{Model 3:}\ \text{logit}(P_{Pos\_test}) = \omega_1 + \beta_1 X_{Gen} + \beta_2 X_{Ele} + \beta_3 X_{Edu} + \beta_4 X_{Emp} + \beta_5 X_{Mar}$$

where P_{Pos_test} denoted the probability of a positive test for HIV/AIDS and $P_{\overline{Pos\_test}}$ is the complement. We fit models #1, #2 and #3 using SAS.

ORDINAL MULTINOMIAL MODELS WITH BAYES ESTIMATES

DERIVED MULTINOMIAL

We considered information provided by the three binary responses [blood test, awareness of campaign, and knowledge of disease] to jointly model through a multinomial, Agresti (2013). Such a combination usually results in a multinomial with $2^3 = 8$ cells. However, we realized empty cells and cells with very few observations. This resulted in a depleted multinomial distribution with four cells. Moreover, the uniqueness of the remaining these 4 cells allowed us to model the data with an ordinal cumulative logistic model with Bayes estimates. The four non-zero combined cells were identified to lie on an ordinal scale based on (0, 0, 0), (0, 0, 1), (0, 1, 1) to (1, 1, 1), referring to (blood test, awareness of campaign, knowledge of disease). We found the scale very useful and interpretable as residents who are aware of HIV/AIDS have an increased risk of testing positive, (Lipsitz, Fitzmaurice, Ibrahim, Sinha, Parzen, M.and Lipshultz, (2009).

We employed an ordinal logistic regression model to determine if gender, education, electricity, marital status and occupation impact responses to blood testing, awareness of the HIV/AIDS campaign, and knowledge of HIV/AIDS simultaneously. The increased number of subpopulations simultaneously allows the researcher to have more insight view of the differential measure. There are 32 subpopulations based on gender, electricity, marital status, education and employment status.

There is Y_{000} =1910 with cell probability p_{000}, denotes those who did not test positive and had no knowledge of HIV/AIDS and had no awareness of a campaign. There is Y_{001} =1814 with probability p_{001} (denotes those who did not test positive and had no awareness of a campaign and had knowledge of HIV/AIDS). There is $Y_{011} = 4341$ with probability p_{011} denotes those who did not test positive and had knowledge of HIV/AIDS

and had awareness of a campaign. There are Y_{111} with probability p_{111} denotes those with a positive test, knowledge of HIV/AIDS and awareness of a campaign, Table 1. We assume $p_{000}+p_{001}+p_{011}+p_{111}=1$.

TABLE 1 ABOUT HERE

A cumulative logit model is

$$\text{logit}[P_i] = \log\left[\frac{P_i}{P_{i-1}}\right] = \theta_j + \sum_{i=1}^{k} \beta_i X_i$$

where X_i denotes the ith covariate and β_i $i = 1, \dots, k$; are the corresponding regression coefficient with θ_j the intercept that goes from $j = 1, \dots, c-1$; where c denotes the number of categories in the multinomial and $P_i \in (p_{000}, p_{001}, p_{011}, p_{111})$ is an ordered set of probabilities. Thus, the cumulative logit model differ by the impact of θ_j.

BAYESIAN ESTIMATES

It is possible that the researcher may have some knowledge of the regression parameters β_i, $i = 1,2,..k$; prior to obtaining a sample. This information often comes from prior surveys and relayed in the form of a distribution. Thus, we used information from Lang and Wilson (2020) about 2009 Mozambique study. We use normal distribution to relay such information on the regression parameter. The depleted multinomial with normal priors on the regression parameters resulted in a posterior distribution. The prior distribution acts as a weighting of the data obtained in the survey through the likelihood [Christensen, Johnson, Branscum, Hanson 2010]. However, there is usually no closed form expression for the resulting posterior, and such was the case with this posterior. We relied on an iterative process to obtain information from the posterior distribution. This principle is handled through a Markov Chain Monte Carlo (MCMC) [Tierney 1994].

This resulted in

$$\text{logit}\,(P_{001}) = \log\left(\frac{P_{001}}{P_{000}}\right) = \theta_1 + \beta_1 X_{Gen} + \beta_2 X_{Ele} + \beta_3 X_{Edu} + \beta_4 X_{Emp} + \beta_5 X_{Mar}$$

where $(\theta_1, \beta_1, \ldots\ldots \beta_5)$has as prior distribution that of the normal distribution. Similarly for the logit (P_{011}), and the logit (P_{111}) . We fit these cumulative multinomial models with Bayes estimates using PROC MCMC.

RESULTS

There are 8,834 Mozambique respondents in the database. Table 2 displays the descriptive statistics for the study. There are 8.7% who had a positive blood test. There are 78.4 % respondents who had knowledge of HIV, and 57.8% who heard of an HIV campaign. Most of the respondents were females, n = 5024 (56.9%) and 58.6 % were either married or lived with a partner. Seventy-four percent of the data did not have electricity at home, 50.2 % had less than 3 years of education. Most of the respondents were employed, n = 6323 (71.6%).

TABLE 2 ABOUT HERE

There were empty cells for the joint outcomes of positive blood test and no knowledge of HIV/AIDS, Table 3. The sample size was also zero for those who tested positive for HIV and were unaware of the disease. In addition, the percentage of respondents who had no knowledge of HIV/AIDS and were aware of an HIV/AIDS campaign were very small.

TABLE 3 ABOUT HERE

The degree of dependency among the three binary responses was also assessed, and the results are shown in Table 4. There is a strong association between knowledge of HIV/AIDS and awareness of an HIV/AIDS campaign ($\Phi = 0.6152$, $p < 0.001$). When blood test and awareness of an HIV/AIDS campaign were analyzed, the correlation weak and yet statistically significant, ($\Phi = 0.1622$). The correlation between blood test and knowledge of HIV/AIDS disease is also statistically significant, but weak as well.

TABLE 4 ABOUT HERE

POLYTOMOUS RESPONSES THROUGH SEQUENCE OF MODELS

The present data structure is unique but applicable for the fit of sequencing models. As such, we fit a sequence of logit models. We began with logit of probability of knowledgeable of HIV/AIDS with key predictors of gender, electricity, education, single and employment. We found that those residents having electricity, who were educated, and who were married were more likely to be knowledgeable of HIV/AIDS, Table 5.

TABLE 5 ABOUT HERE

In Stage #2, we fit a logit of probability of HIV/AIDS awareness campaign. Of those who were knowledgeable, 5,110 residents heard about the campaign and 1,814 did not. In stage #2, we fit a model only to the residents who were knowledgeable of the campaign. We found that males, having electricity in the home, educated, and marital status were key factors in being aware of the campaign, Table 5.

In Stage #3, we fit the logit of testing positive. Seven hundred and sixty-nine residents tested positive and 4,341 tested negative. In stage #3, we fit a model only to

253 the residents who were knowledgeable of the campaign. We found that of those aware
254 and knowledgeable, females and employed were more likely to test positive, Table 5.
255 The sequence of binary model is particularly useful when one wants to investigate
256 certain subgroups. Education is a key factor in people being knowledgeable about AIDS
257 and being aware of the campaign.

258 **MULTINOMIAL LOGISTIC REGRESSION WITH BAYES ESTIMATES**

259 The data structure with empty cells and limited information for certain response
260 categories necessitated the need to find a model that is less affected by the extremity in
261 the distributions and in certain subgroups, Albert and Chib (1993). As such, we fit a
262 multinomial ordinal logistic regression model with Bayesian estimates, where

263 $$\log\left(\frac{P_{111}}{P_{011}}\right) = \beta_0 + \beta_1 Educ_i + \beta_2 Elec_i + \beta_3 Sing_i + \beta_4 Gender_i + \beta_5 Work_i$$

264 $$\log\left(\frac{P_{011}}{P_{001}}\right) = \beta_0 + \beta_1 Educ_i + \beta_2 Elec_i + \beta_3 Sing_i + \beta_4 Gender_i + \beta_5 Work_i$$

265 $$\log\left(\frac{P_{001}}{P_{000}}\right) = \beta_0 + \beta_1 Educ_i + \beta_2 Elec_i + \beta_3 Sing_i + \beta_4 Gender_i + \beta_5 Work_i$$

266 with prior probabilities for the regression coefficients β_i $i = 1,..\ 5$; where $Educ_1$ denotes
267 education, $Elec_2$ denotes electricity, $Sing_3$ denotes living alone, $Gender_4$ denotes gender,
268 and $Work_5$ denotes employment with the corresponding β parameters. The posterior
269 distribution was set to obtain samples of N=2500. Table 6 provides a summary of
270 interval estimates for the odds ratios. These data reveal that gender, electricity, being
271 single having a job and education are key contributors overall in society. There is a wide
272 gap of [21.546, 28.534] or [22 to 29 times more likely] between those who have
273 knowledge of HIV/AIDS, heard of campaign, and test positive versus those who have
274 knowledge of HIV/AIDS, heard about the campaign and tested negative. There were
275 more residents wo had tested negative, had no awareness of the campaign but had
276 knowledge of HIV/AIDS that those who had tested negative but had no knowledge of the
277 campaign or HIV/AIDS, [OR 0.406, 0.363]. It seems that while knowledge of HIV/AIDS

is key there are other factors that need to be examined in the constant fight against the disease.

TABLE 6 ABOUT HERE

The diagnostic plots, trace plot, autocorrelation plot and kernel density plot indicate that all posterior estimates converged. An example of those plots is given in Figure 2. The trace plot is constant over the graph indicates that the Markov chain has stabilized. The autocorrelation implies good mixing. Figure 2 suggests that no high degree of autocorrelation exists for odds ratio at employed of the posterior samples. Similar results were obtained for the other parameters. The kernel density plot, estimates the posterior marginal distribution of the parameters, showing the posterior estimate has the highest density, Figure 2.

FIGURE 2 ABOUT HERE

CONCLUSIONS

It is imperative when analyzing survey data to do so when possible with simultaneous modeling. In this research, we analyzed gender, education, electricity, employment and marital status in addressing the impact on blood test (measures of risk), awareness, and knowledge of HIV/AIDS, simultaneously. The data structure was not the usual normal data patterns. We encountered empty cells and cells with too little information to fit the usual statistical model. The empty cells and the little information led to a unique set of data. On one hand, it allowed a set of sequences of submodels. On the other hand, it leads to an ordering of the cells, which was informational. We used two related but different models that looked at the responses simultaneously.

We found that gender, education, electricity, employment and marital status were key factors in addressing these measures of risk, awareness, and knowledge jointly.

However, in looking at polytomous models (conditional), we found that education, while important, is not a factor if one had knowledge of HIV/AIDS and awareness of the campaign. This may be due to the fact that knowledge and awareness are more prevalent as they are educated.

The multinomial model with Bayesian estimates (marginal) tells, overall, the single women were likely testing negative, aware of the campaign, and knowledgeable about HIV/AIDS. Women who are not single and have electricity in the home are more likely to test positive. Those who have electricity, employed and have education are more likely to test positive, unaware of the campaign, and have no knowledge of HIV/AIDS.

This paper discussed two alternative approaches but both addressed simultaneous fit. The benefits afforded by any of these approaches are significant. The two models addressed different questions. The polytomous model tells us about conditional mean modeling. While the multinomial model gave us results about marginal mean modeling.

Abbreviations

HIV - human immunodeficiency virus

AIDS - Acquired Immunodeficiency Syndrome

Declarations

- **Ethics approval and consent to participate**

  Not needed

- **Consent for publication**

  Not needed

- **Availability of data and materials**

Attached

- **Competing interests**

There is no competing interests

- **Funding**

Neither Dr. Dornelles, Dr. Fang, Dr. Wang nor Dr. Wilson had any funding in regards to this research. There are no funders.

- **Authors' contributions**

The four authors equally contributed. Research and Background- Wang and Dornelles; Model and Computing Fang and Wilson; Findings and Conclusions: Dornelles, Fang, Wang and Wilson. Editing the manuscript Dornelles, Fang, Wang and Wilson

- **Acknowledgements**

No

- **Authors' information (optional)**

Dr. Adriana Dornelles is a biostatistician at Arizona State University

Dr. Di Fang is a professor of health economics in the Department of Agribusiness at University of Arkansas

Dr. Ying Wang is a research specialist at Banner Health.

Dr. Jeffrey Wilson is a professor of Statistics at Arizona State University

References

Agresti A (2013) Categorical Data Analysis - 3rd edition John Wiley & Sons, Inc

Albert J, Chib S. Bayesian inference for logistic regression parameters. Journal of the American Statistical Association. 1993;88:669-79.

Audet CM, Burlison J, Moon TD, Sidat M, Vergara AE, Vermund SH. Sociocultural and epidemiological aspects of HIV/AIDS in Mozambique. BMC Int Health Hum Rights. 2010;10:15. doi: 10.1186/1472-698X-10-15. PubMed PMID: 20529358; PubMed Central PMCID: PMCPMC2891693.

Christensen R, Johnson W, Branscum A, Hanson T. Bayesian ideas and data analysis: An introduction for scientists and statisticians. CRC Press; 2010.

Efron B (2015) Frequentist accuracy of Bayesian estimates Volume 77, Issue 3 Pages: 509-726 JRRS Series B

Font A, Puigpinos R, Chichango IE, Cabrero N, Borrell C. AIDS-related knowledge and behaviors in Mozambique. Rev Epidemiol Sante Publique. 2006;54(4):341-54. PubMed PMID: 17088698.

Lang A (2019) Differences in Modeling Simultaneous to Separate response in Hierarchical model: Applications in Mozambique survey Data *Master Paper* ASU

Leathart, T., Pfahringer, B., Frank, E. (2016). Building ensembles of adaptive nested dichotomies with random-pair selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 179–194). Berlin: Springer.

Lipsitz, SR, Fitzmaurice, GM, Ibrahim, JG, Sinha, D. Parzen, M.and Lipshultz, S. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: An application to acquired immune deficiency syndrome data *J. R. Statist. Soc.* A (2009) 172, Part 1, pp. 3–20

371 Risk analysis. HIV / AIDS country profile: Mozambique. AIDS Anal Afr. 1996;6(6):8-
372 11. PubMed PMID: 12320532.
373 Tierney L. Markov chains for exploring posterior distributions. The Annals of
374 Statistics. 1994;22:1701-28.
375

Figures

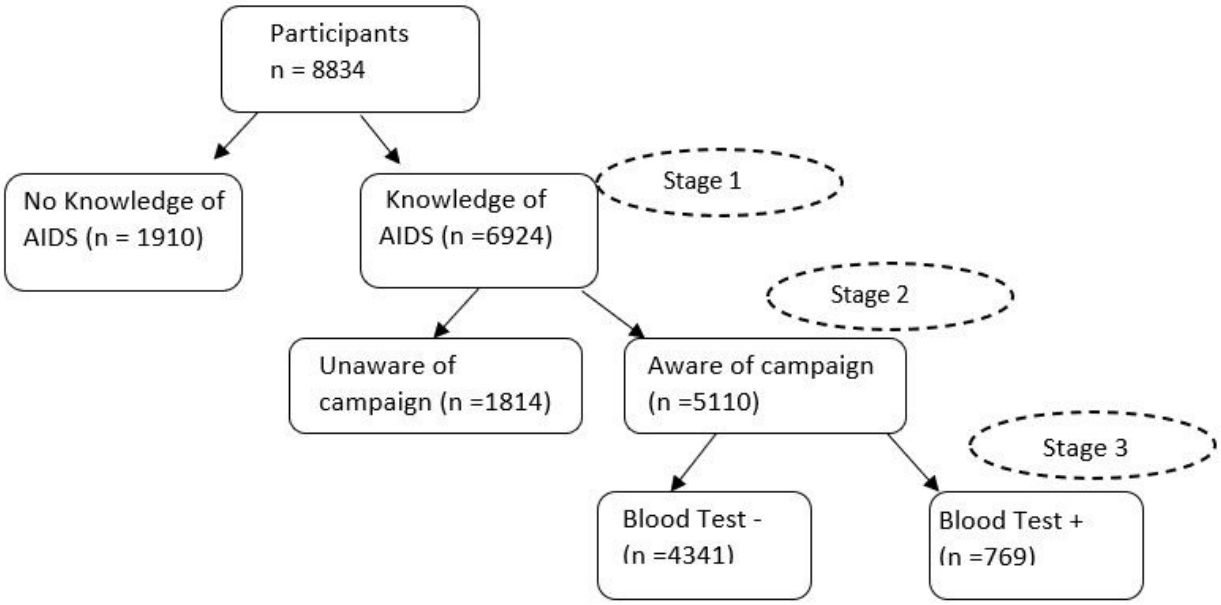

Figure 1

NESTED DICHOTOMIES TREE FOR THE MOZAMBIQUE HIV/AIDS SURVEY.

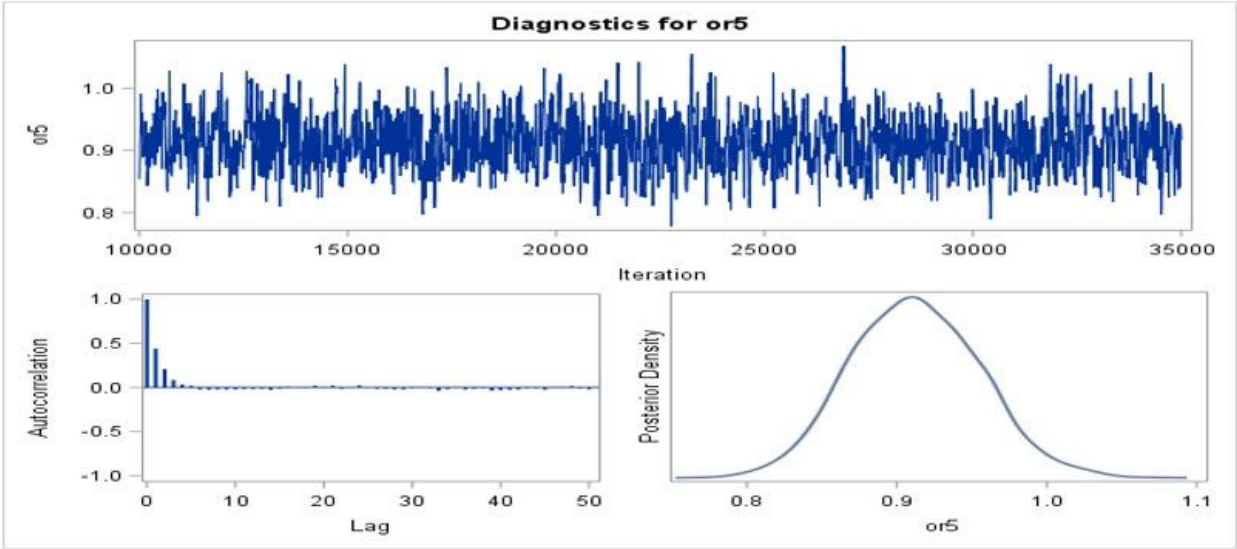

Figure 2

DIAGNOSTIC GRAPHS FOR EMPLOYMENT (WORK)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixA.docx