STA 471 2.0

Generalized Linear Model Project

R.H.T.O.Wickramarathne

AS2018581

Contents

List of Tables

List of Figures

Background and Introduction

The acquired immune deficiency syndrome (AIDS) and the human immunodeficiency virus (HIV) are fatal illnesses that cannot be cured.

HIV (human immunodeficiency virus) is a virus that targets immune cells, rendering a person more susceptible to other infections and disorders. If HIV is not treated, it can develop the disease AIDS (acquired immunodeficiency syndrome). AIDS is the advanced stage of HIV infection that happens when the virus severely compromises the body's immune system.

At the end of 2018, there were 37.9 million HIV-positive persons living in the world, according to UNAIDS. There were 1.7 million kids and 36.2 million adults among them. In 2018, approximately 79% of HIV-positive individuals knew their status. About 8.1 million people, or the remaining 21 percent, still require access to HIV testing services in order to determine their HIV status. A key entry point for HIV prevention, care, treatment, and support services is HIV testing. The majority of HIV patients live in low- and middle-income nations. Eastern and southern Africa had 20.6 million (57%) HIV-positive individuals in 2018, while Asia and the Pacific had 5.9 million (16%), western and central Africa had 5.0 million (13%) and North America had 2.2 million (6%).

According to the final report of the Sri Lanka Demographic and Health Survey in 2016, since the first HIV-infected Sri Lankan was identified in 1987, a total of 2,308 HIV-positive people has been recorded up to the end of 2015. In 2015, 235 HIV cases were reported to the National STD/AIDS Control Program (NSACP), which organizes, plans, and implements the country's HIV National Strategic Plan and AIDS Policy. However, the stated figures represent only a proportion of the HIV-infected people in the country. Many infected people may be unaware of their HIV status, and stigma and discrimination toward HIV-affected people reduce voluntary HIV testing.

Data description

The Department of Census and Statistics (DCS) conducted the Sri Lanka Demographic and Health Survey (DHS) in 2016. The survey used Computer-Assisted Personal Interviewing (CAPI) with mobile and wireless technology to collect data. For the first time in DCS history, data entry and validation of DHS 2016 were also done on-site at CAPI using the digital questionnaire on tablet computers. A total of 28,720 housing units were chosen for the sample, of which 27,455 were occupied at the time of the survey. Of the 27,210 successful interviews, 27,455 were from existing households.

This study is based on a single myth about HIV/AIDS in married women. The Myth is that "people can obtain HIV through mosquito bites," and three categories were used to assess the responses: "Right," "Wrong," and "Don't know." 18302 ever-married women were sampled, and their knowledge about the above Myth and variables connected to their socio-demographic characteristics are considered for this study.

Table 1: Data description

Variable	Description
Residence	1 = Urban 2 = Rural 3 = Estate
Region	1 = Western 2 = Central 3 = Southern 4 = Northern 5 = Eastern 6 = North Eastern 7 = North Central 8 = Uva 9 = Sabaragamuwa
Religion	1 = Buddhist 2 = Hindu 3 = Islam 4 = Roman Catholic 5 = Other Christian 6 = Other
Ethnicity	1 = Sinhala 2 = Sri Lanka Tamil 3 = Indian Tamil 4 = Sri Lanka Moor/ Muslim 5 = Malay 6 = Burger 7 = other
Age group	1 = 15-19 age group 2 = 20-24 age group 3 = 25-29 age group 4 = 30-34 age group 5 = 35-39 age group 6 = 40-44 age group 7 = 45-49 age group
Marital Status	1 = Married or Living together 2 = Divorced/Separated 3 = Widowed 4 = Never married/Never Lived Together 5 = Married but never lived together
Current marital status	1 = Currently married 2 = Living with a man 3 = Not in union/Husband died/Divorced/Separated
Wealth index	1 = Lowest 2 = Second 3 = Middle 4 = Fourth 5 = Highest
Highest Educational Qualification	1 = No education (77&88) 2 = Passed Grade 1-5 3 = Passed Grade 6-10 4 = Passed G.C.E.(O/L) or equivalent 5 = Passed G.C.E.(A/L) or equivalent 6 = Degree and above

Frequency of Radio Listening	1 = At least once a week 2 = Less than once a week 3 = Not at all
Frequency of reading Newspapers/Magazines	1 = At least once a week 2 = Less than once a week 3 = Not at all/Cannot read
Frequency of Watching Television	1 = At least once a week 2 = Less than once a week 3 = Not at all
Frequency of all media combined	1 = At least once a week 2 = Less than once a week 3 = Not at all
Have you ever given birth	1 = Yes 2 = No
Working Status	1 = Yes 2 = No
Current pregnancy status	1 = Yes 2 = No 3 = Don't know
People can get HIV virus from mosquito bites(Y1)	0 = Right 1 = Wrong

Objectives

To learn how ever-married women's socio-demographic variables influence their understanding of the above HIV/AIDS myth.

Methodology

This study is based on a single myth about HIV/AIDS in married women. This project aims to investigate the knowledge regarding the Myth of HIV/AIDS that varies on ever-married women's socio-demographic characteristics. This project's implementation phase has been investigated in three major sections. The first section conducts exploratory data analysis. The second section performs data analysis and predictive model and fits the model, and the third section verifies the assumptions of the logistic model and then validates the model. The data was loaded and processed using R software.

The data set contains 15,121 observations and 16 independent variables. All 16 variables were categorical. The dependent variable is based on the Myth, which is "People can get HIV virus from mosquito bites-Y1," and three categories have been used to measure respondent's outcomes: "Right", "Wrong," and "Don't know".However, in this case, we consider the individuals whose outcomes are "Right" and "Wrong".

As the most common way of displaying qualitative data, bar graphs were used to explore each qualitative variable in this study. Bar charts were used to compare categories with at least one categorical or discrete variable. Each bar represents a summary value for one discrete level, where longer bars indicate higher values. Moreover, cluster bar charts were used to explore two

qualitative variables. The Chi-square independence test was used to check whether Y1 and each predictor variable are likely to be related or not. The hypothesis to be tested for the Chi-square independence test is as follows,

Ho: The two variables are independent (Y1 and Socio-demographic variable)

H1: The two variables relate to each other.

Decision rule : Reject Ho if p-value < 0.05 (α)

Before the data analysis, the whole data set was split into two parts train set and a test set.80% of observations were used as a training test, while the remaining data were used as a test set. Logistic regression has been used to estimate the relationship between a dependent variable and independent variables in this case since there is a binary outcome for the dependent variable, such as right or wrong. The backward elimination method was applied to choose the best-fitted model. This process begins with a full model and, at each step, gradually eliminates variables from the full model to find a reduced best model that best explains the data. AIC has been used to eliminate data from each step. AIC is known as Akaike Information Criterion, a method for selecting a model. In this study, we generate models and eliminate variables with the lowest AIC value in each model. Then finally, we got the best model with the lowest AIC value (Zhang, 2016a)

The AIC statistic is defined for logistic regression as follows,

- AIC = -2/N * LL + 2 * k/N

  Where *N* is the number of examples in the training dataset, *LL* is the log-likelihood of the model on the training dataset, and *k* is the number of parameters in the model.

- First, we fit a model using all p predictors. Define this as Full model,M_P.
- Next, for k = p, p-1, … 1, we fit all k models that contain all but one of the predictors in M_k ,for a total of k-1 predictor variables. Next, pick the best among these k models and call it M_{k-1}.
- Lastly, we pick a single best model from among $M_0 \ldots M_p$ using AIC.

After choosing the model, check whether the basic assumptions of the logistic regression are satisfied. In our study, we check the following assumptions. Independence of errors, presence of outliers and lack of strongly influential cases.(“Assumptions of Logistic Regression,” n.d.) Standardized residual plot used to show independence of errors and absence of outliers.

Cook's distance plot is used to detect the influential cases since Cook’s distance is a summary measure of influence cases. A significant value of Cook's distance indicates an influential observation. Most influential cases can be identified by exploring Cook's distance plot and examining the change of coefficients after identifying and removing these influential observations. If coefficients are changed minimally, those observations can be counted as not influential(Zhang, 2016b) Also, we can compare the accuracy of the first model and the model without influential cases. If accuracy is approximately the same, then the first model chooses our best model.

Model validating is done as a final section of data analysis. It was done by using test data set. Hosmer and Lemeshow's goodness of fit (GOF) test was applied to the test data set. Following hypothesis used to be tested,

Ho: The model is adequate

H1:Model is not adequate

Decision rule : Reject H0 if p-value < 0.05 (α)

If Ho is rejected, we should add additional terms to the model.

Exploratory Data Analysis

Exploring one qualitative variable

Distribution of ever-married women

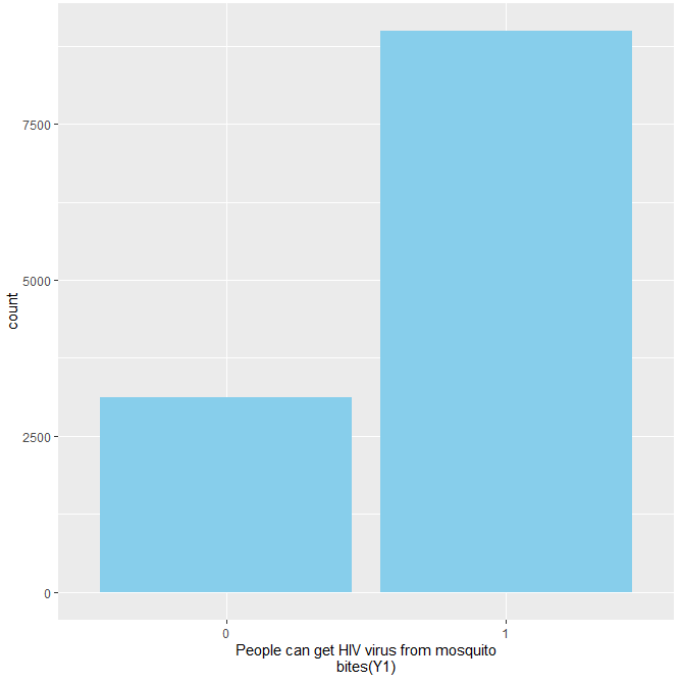

Figure 1:Bar chart of distribution of ever-married women

According to figure 1, it represents most of the ever-married women do not accept the Myth that people can get HIV from a mosquito.

Distribution of ever-married women by Residence

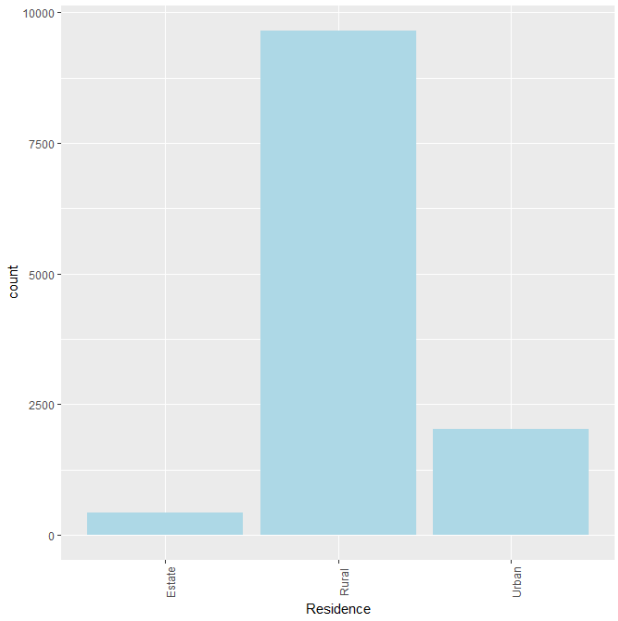

Figure 2:Bar chart of distribution of ever-married women by Residence

Figure 2 compares the counts of types of Residence where respondents live. The rural area is the most common, followed by estate and urban.

Distribution of ever-married women by Religion and Ethnicity

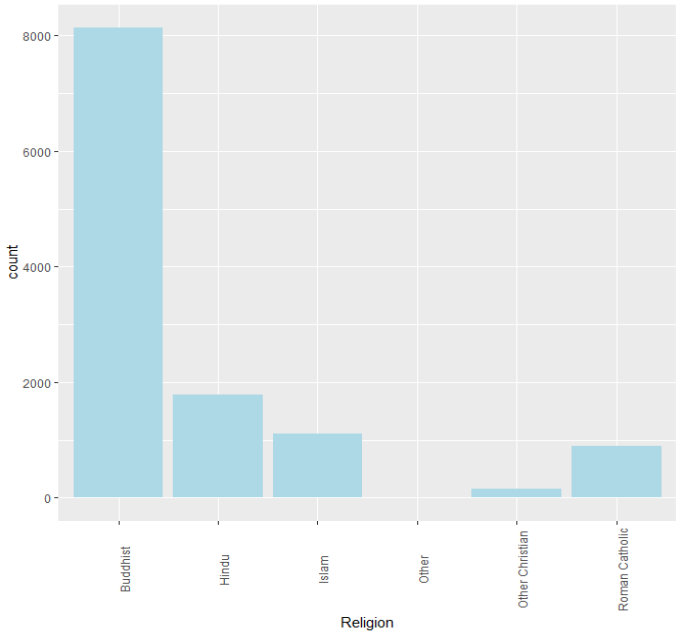

Figure 3:Bar chart of distribution of ever-married women by Religion

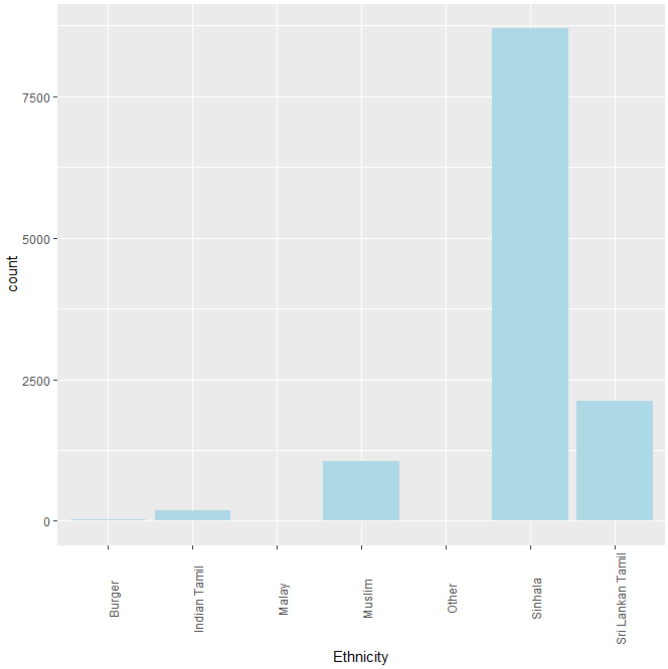

Figure 4:Bar chart of distribution of ever-married women by Ethnicity

Figure 3 represents that most of the women believe in Buddhism compared to other religions and figure 4 represents that most of the women are Sinhalese compared to other ethnic groups.

Distribution of ever-married women by Age group

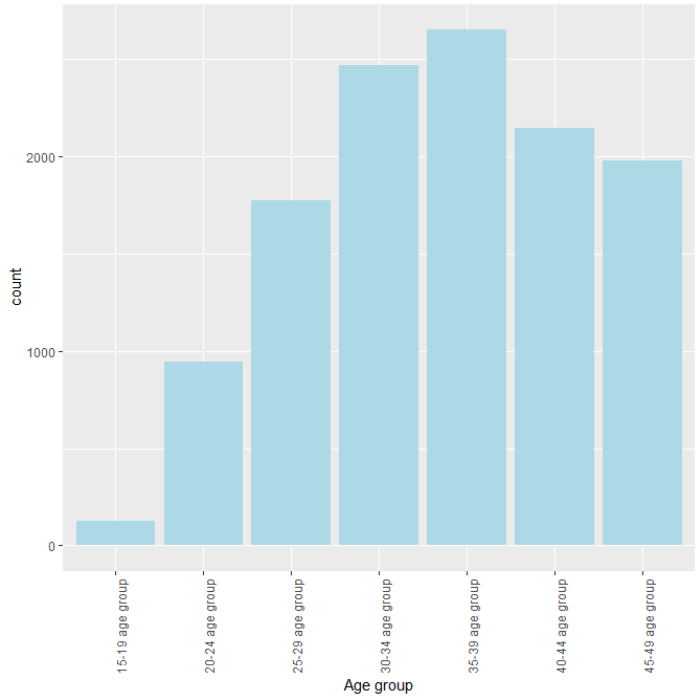

Figure 5:Bar chart of distribution of ever-married women by age group

Figure 5 shows that most of the women are included in the 35-39 age group. Out of the respondent least number of respondents were included in the 15-19 age group.

Distribution of ever-married women by Marital status and Current marital status

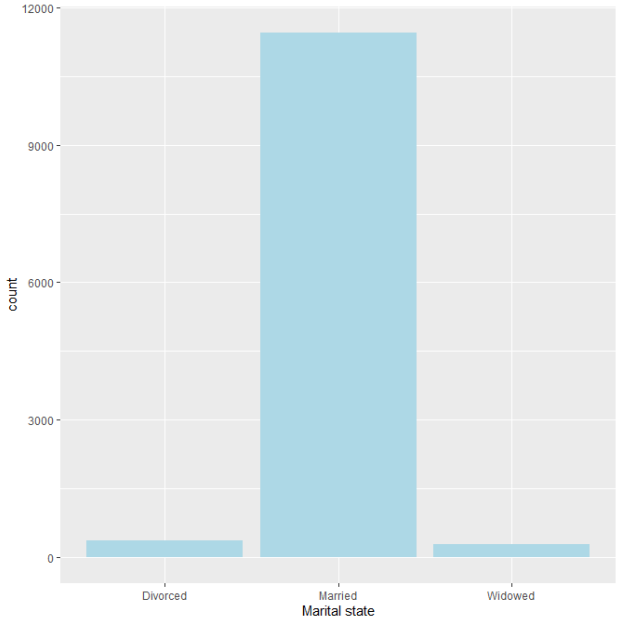

Figure 7:Bar chart of distribution of ever-married women by marital status

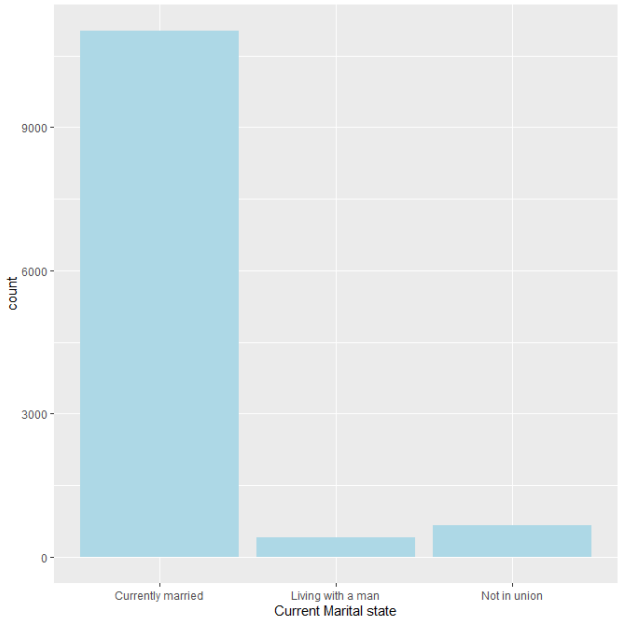

Figure 6:Bar chart of distribution of ever-married women by current marital status

According to figure 5, most women are married. Also, figure 7 shows that their current marital status is also married compared to other states. Moreover, most minor women live with a man without marrying them.

Distribution of ever-married women by Highest Education Qualification

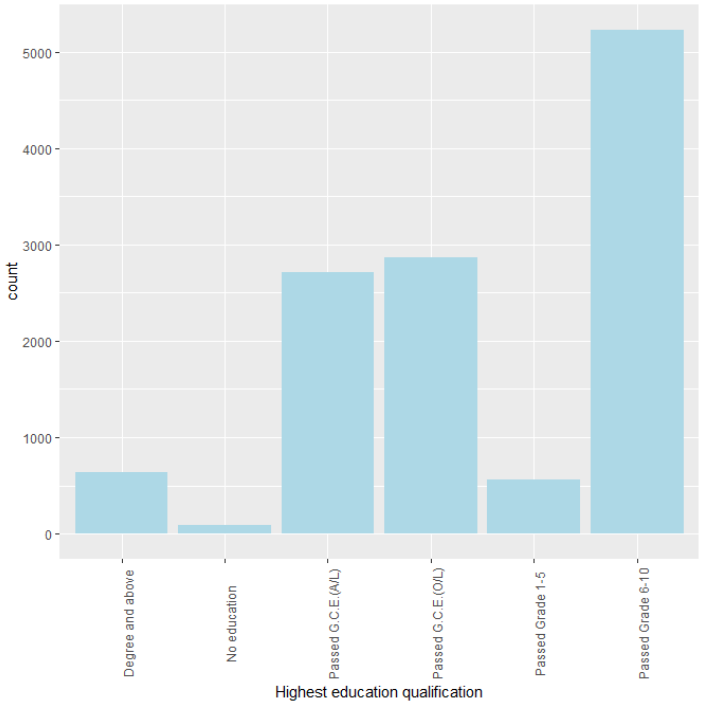

Figure 8:Bar chart of distribution of ever-married women by highest education qualification

Figure 8 shows that most of the respondents passed grades 6-10 compared to other groups. According to the bar chart, few respondents did not educate.

Distribution of ever-married women by Frequency of reading newspaper and Watching television

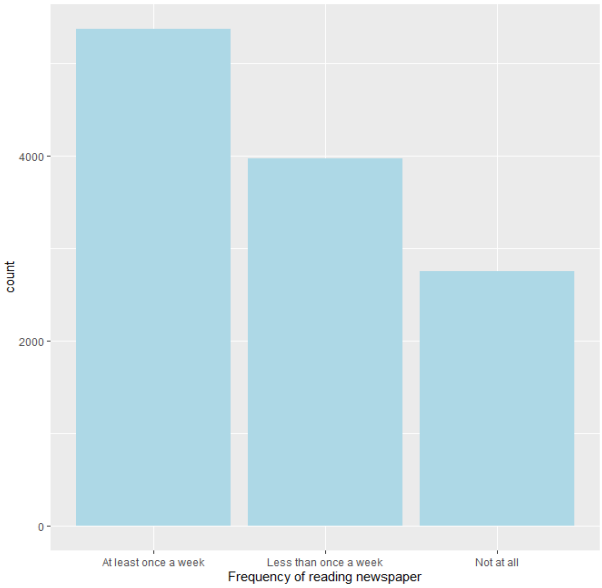

Figure 9:Bar chart of distribution of ever-married women by Frequency of reading newspapers

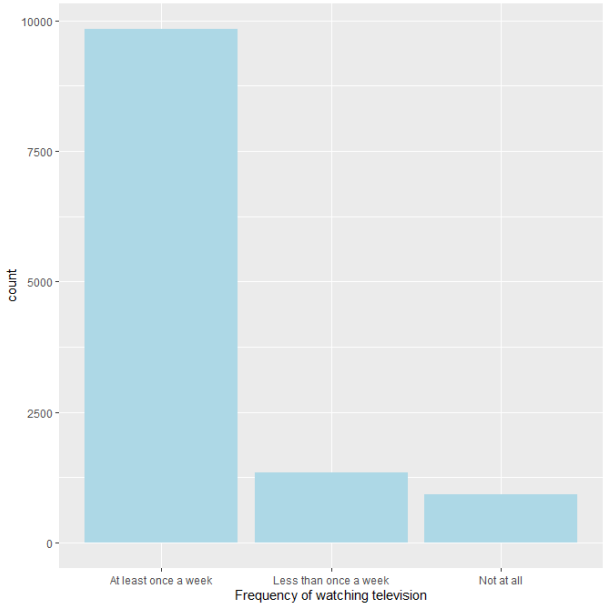

Figure 10:Bar chart of distribution of ever-married women by Frequency of watching television

Distribution of ever-married women by Frequency of listening to the radio

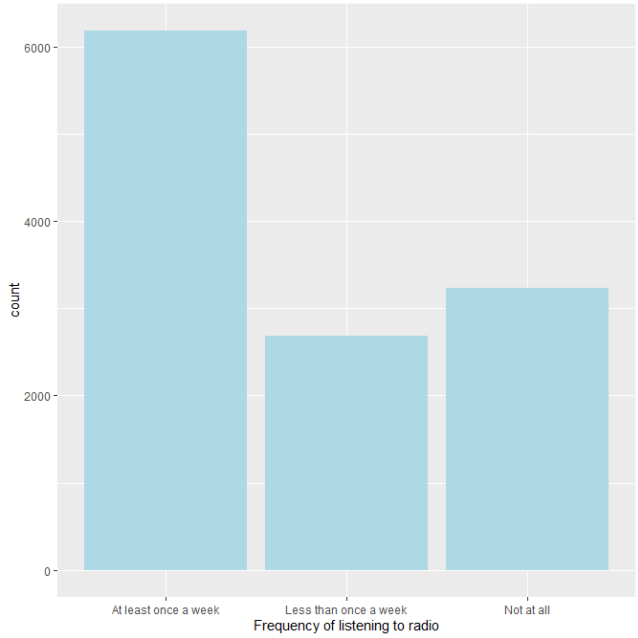

Figure 11:Bar chart of distribution of ever-married women by Frequency of listening to the radio

Figures 9,10 and 11 show that most women listen to the radio, watch television and read newspapers at least once a week. However, some respondents neither do anything mentioned above.

Distribution of ever-married women by Frequency of all media combined.

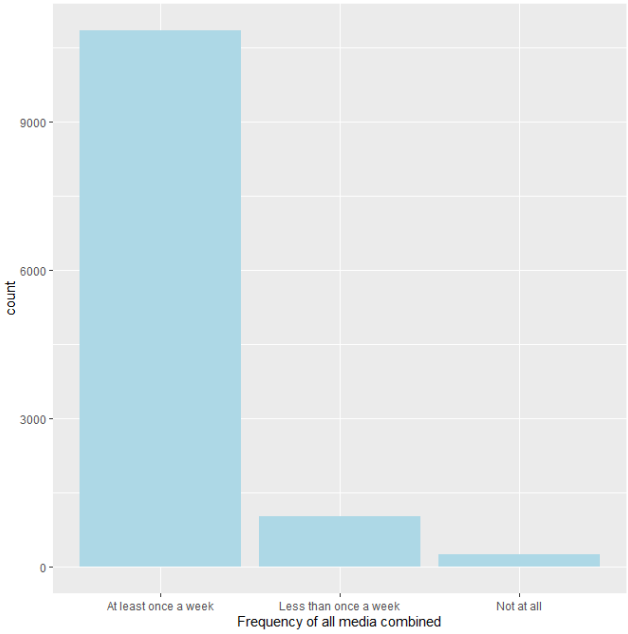

Figure 12:Bar chart of Frequency of all media combined.

Figure 12 depicts that the Frequency of respondents engaging with all media is more likely to be at least once a week. However, there are few respondents who did not engage in any of media.

Distribution of ever-married women by the status of given birth ever and current pregnancy status

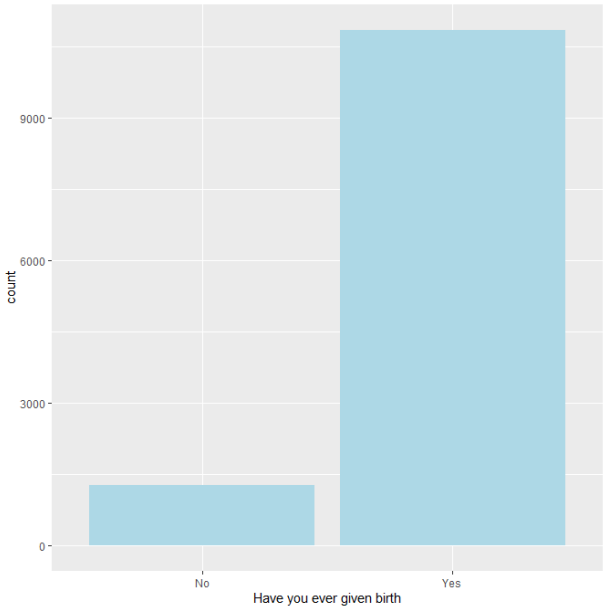

Figure 14:Bar chart of distribution of ever-married women by the status of given birth ever

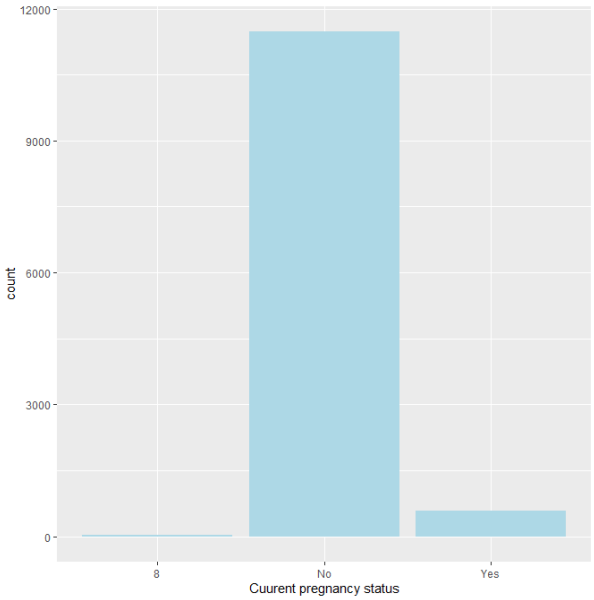

Figure 13: Bar chart of distribution of ever-married women by current pregnancy status

Figure 13 illustrates that most women who participated in this study did give birth to a child and fig.14 show the most of respondents' current pregnancy status is not when compared to current pregnancy is yes and do not know.

Distribution of ever-married women by working status

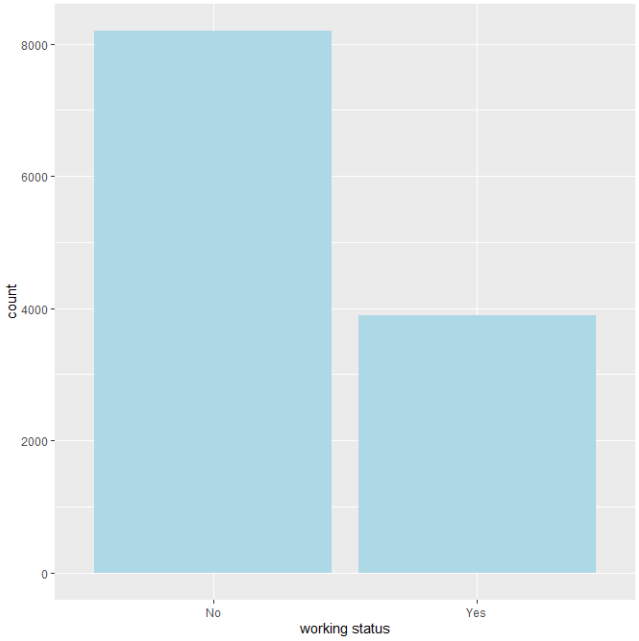

Figure 15:Bar chart of distribution of working status

This bar chart compares the counts of the working status of respondents. Among the respondents, most of the women do not work.

Distribution of ever-married women by wealth index

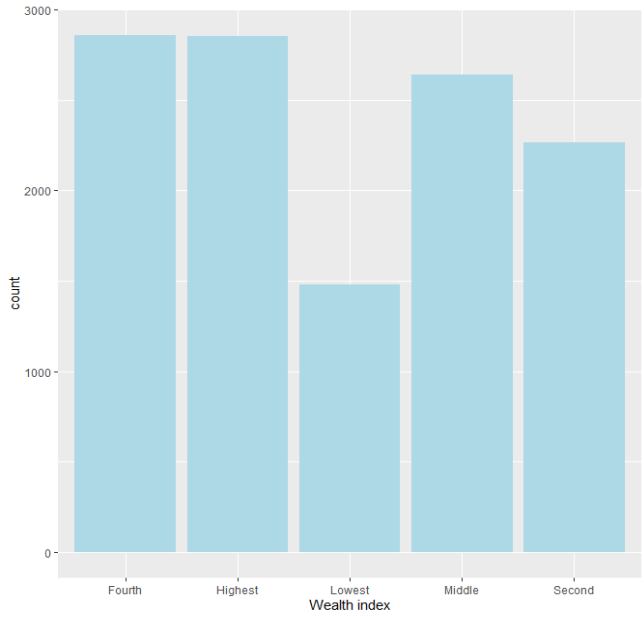

Figure 16:Bar chart of distribution of ever-married women by wealth index

Figure 16 shows that an equal number of women is included in the fourth and highest wealth index. Furthermore, the lowest count of women includes the lowest category.

Exploring two qualitative variables

Distribution of ever-married women by Y1 and Region

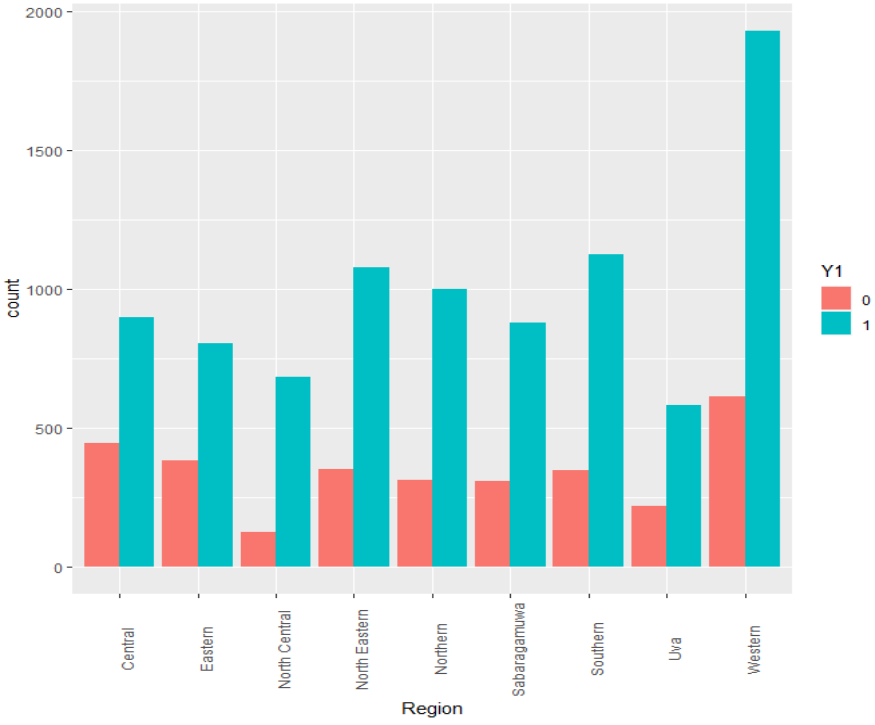

Figure 17:Cluster bar chart of distribution of ever-married women by Y1 and Region

Figure 17 shows that most women do not accept the Myth in every region. Moreover, most of the respondents live in the western province than in other regions.

Distribution of ever-married women by Y1 and Religion

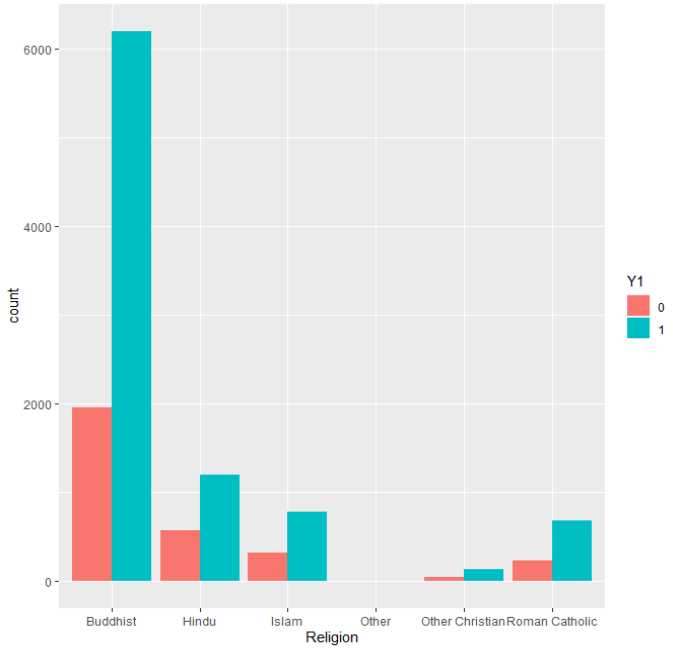

Figure 18:Distribution of ever-married women by Y1 and Religion

According to this graph, most women who do not accept the Myth of HIV, believe in Buddhism. In each region, most women do not accept the Myth rather than accept the Myth.

Distribution of ever-married women by Y1 and Working status

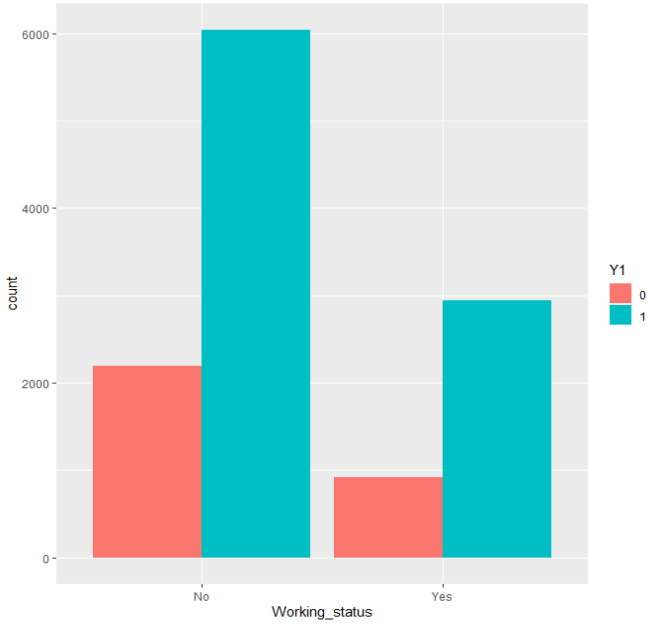

Figure 19:Cluster bar chart of distribution of ever-married women by Y1 and working status

This graph shows that most women who do not accept the Myth do not engage in work.

Distribution of ever-married women by Y1 and Ethnicity

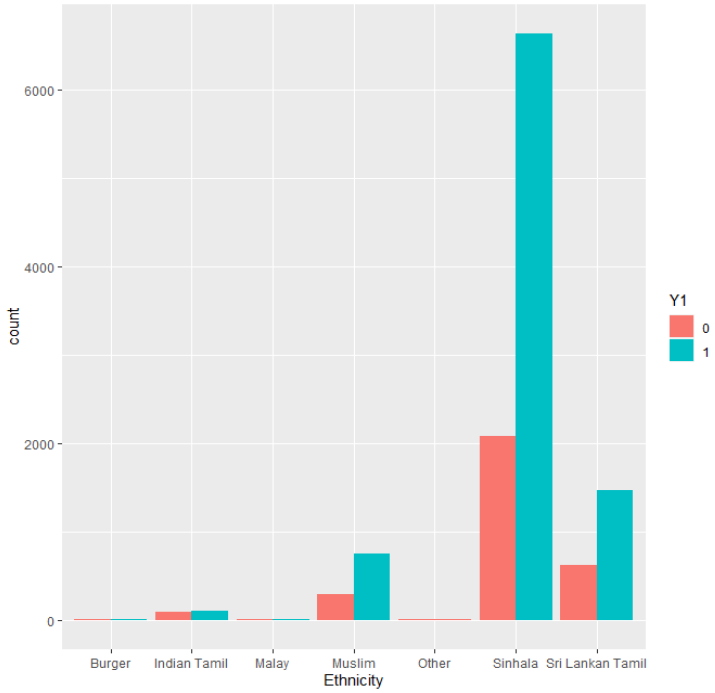

Figure 20:Cluster bar chart of distribution of ever-married women by Y1 and Ethnicity

According to figure 20, most women are Sinhalese, and most respondents do not accept the Myth. Among the Sinhalese respondents, the greatest number of respondents have right idea of the myth.

Distribution of ever-married women by Y1 and Marital status

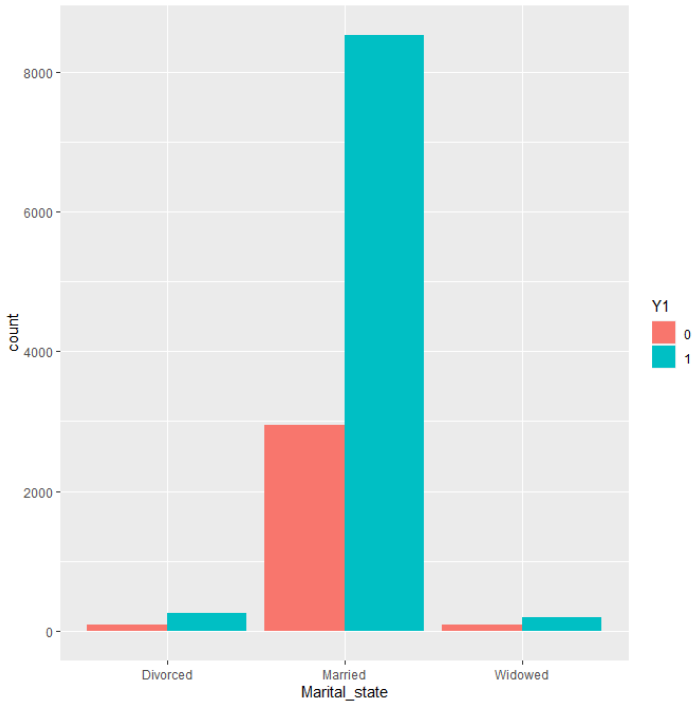

Figure 21:Cluster bar chart of distribution of ever-married women by Y1 and Marital status

Figure 21 shows that most respondents are married compared to other states. Among married women, most women do not accept the Myth.

Distribution of ever-married women by Y1 and Highest education qualification

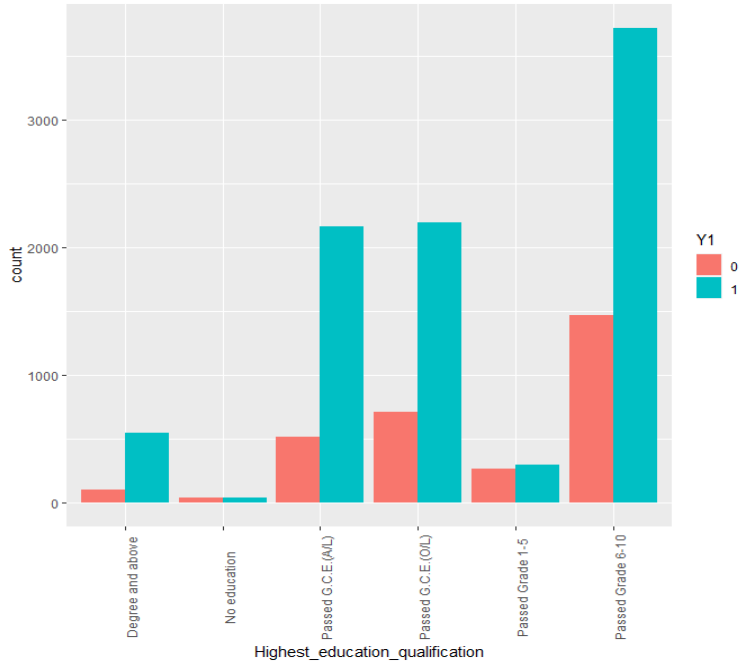

Figure 22:Cluster bar chart of distribution of ever-married women by Y1 and Marital status

Figure 22 indicates that in each category of education qualification, the majority of respondents do not believe the Myth. Furthermore, for no education level, the number of respondents who accept and do not accept the myth is equal.

Distribution of ever-married women by Y1 and Frequency of all media combined.

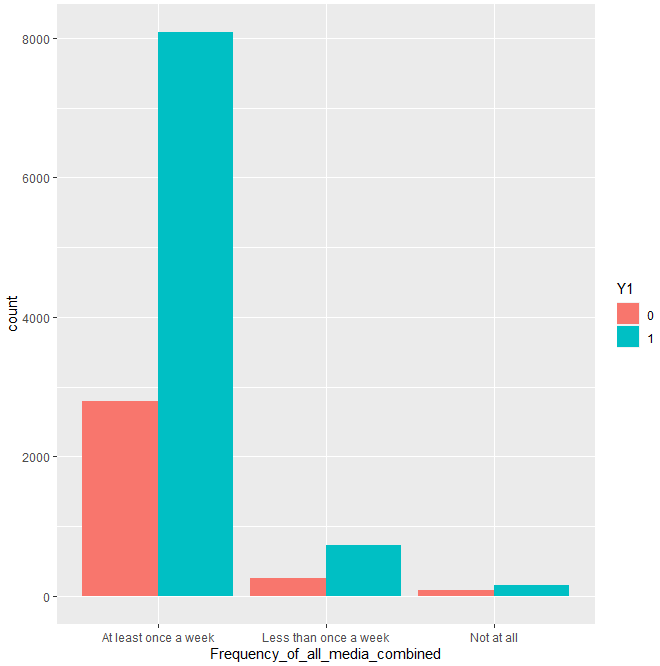

Figure 23:Cluster bar chart of ever-married women by Y1 and Frequency of all media combined.

This graph shows that most of the women engaged in all media at least once a week who do not accept the Myth of the HIV virus.

Distribution of ever-married women by Y1 and Wealth index

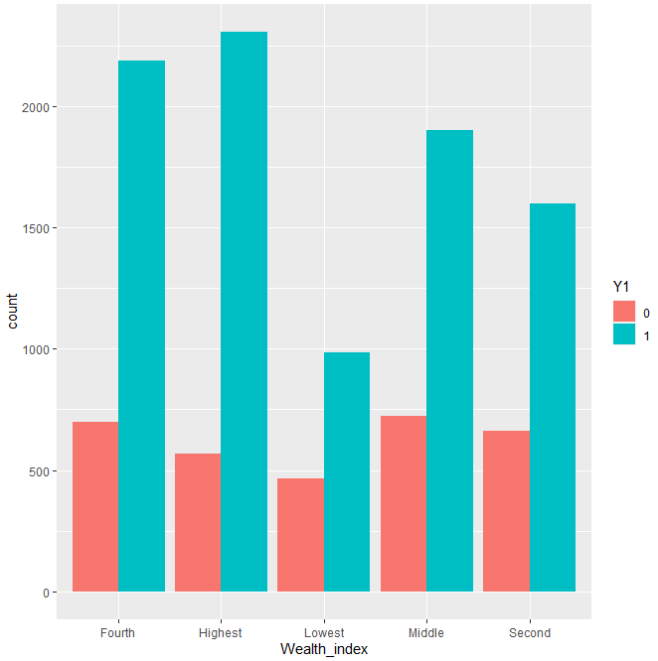

Figure 24:Cluster bar chart of distribution of ever-married women by Y1 and Wealth index

Figure 24 indicates that the majority of women who reject the Myth include each wealth index category.

Chi-square independence test

Hypothesis to be tested,

Ho: The two variables are independent (Y1 and Socio-demographic variable)

H1: The two variables relate to each other

Table 2:Chi-square Independence test

Variable	**X^2**	**P-value**
Y1 vs. Region	143.32	< 2.2e-16
Y1 vs. Religion	77.509	< 2.2e-10
Y1 vs. Ethnicity	115	< 2.2e-16
Y1 vs. Age group	8.7958	0.1854
Y1 vs. Marital status	4.1019	0.1286
Y1 vs. Current marital status	11.997	0.002483
Y1 vs. Highest education qualification	346.42	< 2.2e-16
Y1 vs. Frequency of listening to the radio	18.826	8.165e-05
Y1 vs. Frequency of watching television	8.868	0.01187
Y1 vs. Frequency of reading newspaper	102.21	< 2.2e-16

Y1 vs. Frequency of all media combined	9.6215	0.008142
Y1 vs. Frequency of reading newspaper	102.21	< 2.2e-16
Y1 vs. Wealth index	146.17	< 2.2e-16
Y1 vs. Working status	8.4227	0.003706
Y1 vs. Have you ever given birth	1.8152	0.1779
Y1 vs. Current pregnancy state	0.88123	0.6436

According to table 2, Y1 and other predictors are dependent except “Have you ever given birth” and “Current pregnancy status” predictors. Since their p-values < 0.05, we have enough evidence to reject Ho at a 5% level of significance.

Data Analysis

Since the data set consist of qualitative variables, using dummy variables are essential when building the model. Eight important qualitative variables were founded using the backward elimination method. There are wealth index, Highest Educational Qualification, Region, Ethnicity, Current marital status, Age group, Frequency of reading Newspapers/Magazines and Frequency of Radio listening. Parameter interpretation was based on the reference level.

This study is based on one Myth regarding HIV/AIDS in ever-married women. The Myth is, "People can get HIV virus from mosquito bites", and three categories have been used to measure the respondent's outcomes: "Right", "Wrong," and "Don't know". It was denoted as Y1. However, this study only considered individuals whose outcomes are "Right" or "Wrong".

Before the model fitting data set was divided into two parts, a test set and a train set.80%of observations were used for the train set .other 30% of observations were used as a test set.

Among the 16 qualitative variables, 8 essential variables were chosen. It was done using the Backward elimination method, and Variables were eliminated using AIC (Akaike’s Information Criteria). (Brownlee, 2019)

AIC= -2 (Maximum log-likelihood – Number of parameters in the model)

First, get the full model and eliminate the predictor variable with the lowest AIC value. Then it will be continued until we get the best-fitted model with the most contributive variables.

Following predictor variables have been dropped,

Table 3:Summary of dropped variables

Step	Df	Deviance	Resid, Df	Resid.Dev	AIC
			12043	13300.15	13406.15
Religion	5	4.25276573	12048	13304.40	13400.40
Residence	2	0.02768179	12050	13304.43	13396.43
Frequency of watching television	2	0.35173594	12052	13304.78	13392.78
Current pregnancy status	2	1.13643053	12054	13305.92	13389.92

Have you ever given birth	1	0.22606063	12055	13306.14	13388.14
Marital state	2	2.33387987	12057	13308.48	13386.48
Working status	1	0.98516181	12058	13309.46	13385.46
Frequency of all media combined	2	3.19552153	12060	13312.66	13384.66

Generalized Liner Model

$Y1 = \begin{cases} 0, \text{Right} \\ 1, \text{Wrong} \end{cases}$

$\pi_i = P(Y_i = 1)$

The best fitted main effect model

The best model was selected considering the stepwise backward elimination process with AIC.

The following index was used as the reference level of chosen variables,

- Region1-Western
- Ethnicity1-Sinhala
- Age group1-15-19 age group
- Current marital status1-Currently married
- Education qualification1-Not educate
- Frequency of reading newspapers1-At least one week
- Frequency of radio listening1-At least once a week
- Wealth index1-Lowest

Best fitted model

$\text{Logit}(\pi_i) = \beta_0 + \beta_1$ Region $+ \beta_2$ Ethnicity $+ \beta_3$ Age $+ \beta_4$ Current marital state $+ \beta_5$ Highest education qualification $+\beta_6$ Frequency of reading newspaper $+ \beta_7$ Frequency of listening to radio $+ \beta_8$ Wealth index

Table 4:Summary of model

	Estimate	Std. Error	P value
(Intercept)	-0.36993	0.30303	0.222168
Region-Central	-0.22041	0.07887	0.005199
Region-Southern	0.11231	0.07958	0.158169
Region-Northern	0.52451	0.11100	0.000001
Region-Eastern	0.03259	0.09276	0.725369
Region-North Eastern	0.12713	0.08049	0.114234
Region-North Central	0.72228	0.11029	0.000001
Region-Uva	0.02062	0.09690	0.831512
Region-Sabaragamuwa	0.06825	0.08534	0.423868
Ethnicity-Sri Lanka Tamil	-0.36102	0.08510	0.000001
Ethnicity-Indian Tamil	-0.63033	0.15512	0.000001
Ethnicity-Muslim	-0.14613	0.08164	0.073454
Ethnicity-Malay	-0.65373	0.56325	0.245788

Ethnicity-Burger	-0.69544	0.52052	0.181533
Ethnicity-Other	-13.86366	160.17117	0.931025
20-24 age group	0.38211	0.19465	0.049637
25-29 age group	0.30550	0.18792	0.104011
30-34 age group	0.40693	0.18614	0.028801
35-39 age group	0.41071	0.18575	0.027030
40-44 age group	0.46807	0.18730	0.012452
45-49 age group	0.53075	0.18835	0.004833
Current marital state-Living with a man	0.32713	0.12775	0.010445
Current marital state-Not in the union	0.08788	0.09615	0.360717
Passed grades 1-5	0.16740	0.23904	0.483721
Passed grades 6-10	0.87090	0.23008	0.000154
Passed G.C.E.O/L	1.01797	0.23422	0.000001
Passed G.C.E.A/L	1.26610	0.23734	0.000001
Degree and above	1.47281	0.25899	0.000001
Frequency of reading newspaper-Less than once a week	-0.11618	0.05278	0.027719
Frequency of reading newspaper-Not at all	-0.19492	0.06053	0.001280
Frequency of listening to radio-Less than once a week	0.15678	0.05697	0.005921
Frequency of listening to radio-Not at all	0.07001	0.05222	0.180006
Wealth index- Second	-0.04174	0.07757	0.590519
Wealth index-Middle	-0.02949	0.07992	0.712154
Wealth index-Fourth	0.06078	0.08239	0.460715
Wealth index- Highest	0.21368	0.09135	0.019168

Null deviance: 13798 on 12095 degrees of freedom

Residual deviance: 13313 on 12060 degrees of freedom

Residual Analysis

Figure 25:Standardized residual values

Figure 25 depicts a non-random pattern as well as two groupings. Because we predict a probability for a variable with values of 0 or 1, If the Y1 value is 0, we predict more and the residuals must be negative; if the Y1 value is 1, we underestimate, and the residuals must be positive. This plot demonstrates error independence, with no outliers.

Detecting Influential observations

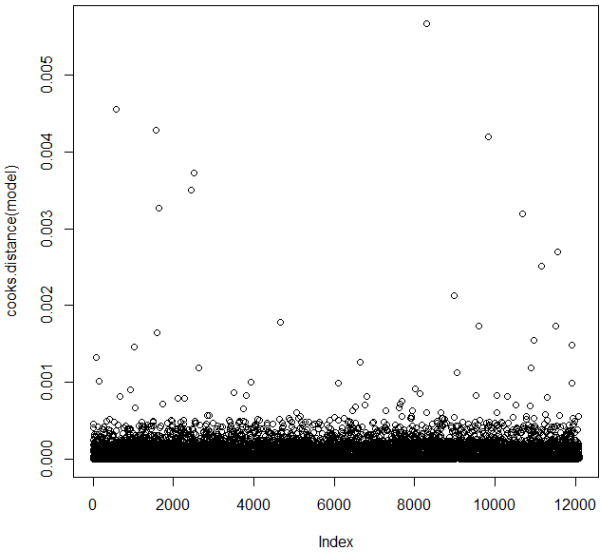

Figure 26:Detecting influential cases

Table 5:Cook's distance of influential cases

Observation	Cook Distance
583	4.550819e-03
2893	5.721351e-04
4325	3.982034e-04
8300	5.669134e-03
8813	1.962570e-07
10667	3.056906e-08

According to the above figure, 583 and 8300 are most likely influential cases. Observation 8300 has the largest value of Cook's distances. Then refit the model by removing influential cases. After that, compare the accuracy of the refit model and the initial model.

Accuracy of refit model=0.7384

Accuracy of initial model=0.7385

According to the output, accuracy has changed minimally after removing influential cases. Therefore, observation 8300 is not influential.

The goodness of fit test

Hosmer and Lemeshow goodness of fit (GOF) test

Hypothesis to be tested,

Ho: Model is adequate

H1:Model is not adequate

X-squared = 8.8431, df = 8, p-value = 0.3557

Since the p-value > 0.05, We do not have enough evidence to reject Ho at a 5% level of significance. Therefore, our fitted model is adequate at a 5% level of significance.

Conclusion

The research project's goal was to fit a general linear regression model that could be used to know how the knowledge regarding the above Myth of HIV/AIDS varies on ever-married women's socio-demographic characteristics. We anticipated a multilinear relationship between the response variable and the predictors selected for this task.

The effects of socio-demographic factors on knowledge regarding the Myth, which is "People can get HIV virus from mosquito bites" were investigated in this research work. The relationship between those factors was studied using a variety of approaches.

In this study, most respondents do not accept the Myth that people can get the HIV virus from mosquitoes.

Out of the respondents in this study, most of women live in rural area and most of them believe in Buddhism. Most of respondents include in the 35-39 age group while the lowest number of respondents include the 15-19 age group. Within the study period most of women were married and they have given birth to child. Moreover, current pregnancy status of women was not. Among the respondent they were passed grade 6-10 while the few respondents did not educate. Also found that most of the respondents were not work ,but most of women are include into the highest and fourth wealth index levels. Also, this study found out most of women reading newspaper ,listening to the radio and watching television at least once week. While few of respondents do not engage any of media.

According to the study, the majority of women who do not accept the HIV myth live in western province. And the majority of them believe Buddhism. Also, there were Sinhalese. This study demonstrated that educated women have a good awareness of myths, and according to the study, the majority of women who passed grades 6-10 do not believe myths, while the number of respondents who accept the myth and do not accept the myth is the same.

According to this study, the majority of married women do not accept the HIV myth. Furthermore, most women do not work, but they are more aware of these myths than working women. Furthermore, respondents who use all media at least once a week are more aware of these fallacies than women who do not use these media. When compared to the other categories, respondents in the highest and fourth wealth indexes do not believe the myth.

As a result, we may conclude that teaching and educating people about these fallacies helps to enhance knowledge and guide them down the proper path. It would also be beneficial in avoiding numerous problems that women may experience in the future. Using new technologies, reading newspapers, watching television, and listening to the radio may provide a clear picture of these misconceptions and treatments.

The data and Chi – squared test suggest that among these 16 variables,Y1 and other predictors are related except “Have you ever given birth” and “Current pregnancy status” predictor variables.

This study's final model was created using a train data set and the backward elimination method. This study made use of logistic regression. It has eight variables. It also has a 73 percent accuracy rate. Build the model again after deleting the influential cases. The accuracy was then about close to 73 percent. As a result, the first model was chosen as the best model.

Standardized residual plot shows the independence of errors and absence of outliers which are assumptions of Logistic regression.Finally, using the test data set, assess the model's goodness of fit. Then it provides an adequate model. As a result, the logistic regression method was appropriate for this type of research. When there are several variables, we can apply backward elimination to find the best model.

References

- Assumptions of Logistic Regression. (n.d.). *Statistics Solutions*. Retrieved July 29, 2022, from http://www.statisticssolutions.com
- Brownlee, J. (2019, October 30). *Probabilistic Model Selection with AIC, BIC, and MDL*. https://machinelearningmastery.com/probabilistic-model-selection-measures/
- Zhang, Z. (2016a). Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine*, *4*(7). https://doi.org/10.21037/ATM.2016.03.35
- Zhang, Z. (2016b). Residuals and regression diagnostics: focusing on logistic regression. *Annals of Translational Medicine*, *4*(10). https://doi.org/10.21037/ATM.2016.03.36