STA 471 2.0 Generalized Linear Models

Project on

How the knowledge regarding a myth of HIV/AIDS varies on ever married women's socio demographic characteristics

Prepared by

AS2018346

Table of Contents

1. Background and Introduction

Human immunodeficiency virus (HIV) attacks the white blood cells in human body. HIV weakens the immune system of human, and it may allow people with HIV more vulnerable to other severe illnesses. Even though HIV does not have a cure, antiretroviral treatment (ART) is used to treat HIV.

First HIV infected person in Sri Lanka was found in 1987. According to the epidemiological HIV data 2020, the estimated number of people living with HIV in Sri Lanka is 3700, however only 2600 knows their status. Moreover 2100 people are living with HIV while receiving ART. The National STD/AIDS Control Programme (NSACP) under the Ministry of health is responsible for supervision of the country's HIV.

There have been problems regarding the spread of HIV by bloodsucking animals like mosquitoes since the beginning of the HIV epidemic. The Centers for Disease Control and Prevention in Atlanta have found through their epidemiological research that there is no proof of spreading HIV from mosquitoes, even in the nations with extraordinary high HIV incidence rates and unchecked mosquito populations.

This study had carried out to find how the knowledge about the myth that the HIV virus can transmit from mosquito bites varies on Sri Lankan ever married women's socio demographic characteristics.

2. Methodology

Objective of the study is to find out how the knowledge regarding the myth “People can get HIV virus from mosquito bites” differ on ever married women’s socio demographic characteristics.

Data set consists of 17 variables and data were collected from 18302 ever married women. The response variable had measured as three categories and for the purpose of the study responses for “Don’t know” cases has been dropped. There were 43 missing values in the variable Frequency of watching television. After removing both don’t know cases and missing values, the sample size reduces to 15121. First exploratory data analysis was done using R software. Since all the variables are qualitative cluster bar charts were drawn for each variable to find the composition of the sample. Significant variables in the full model were taken to find how women have responded to the myth regarding HIV.

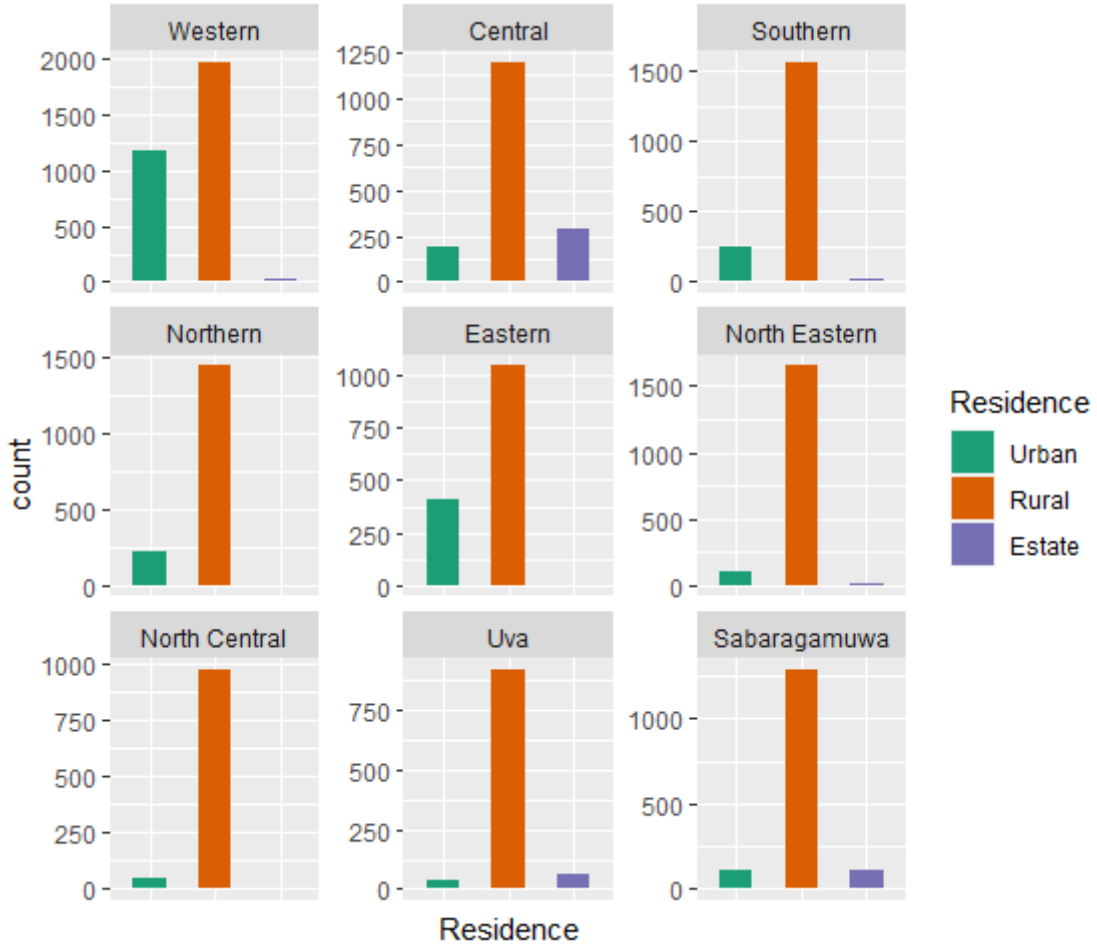
In the next step data set was split as training and testing set. 80% of the data were used as the training set and remaining 20% was used as the testing set. Logistic regression model with a logit link has been fitted to study the objective. Sixteen socio demographic variables have been considered as predictors. Response to the myth has recorded as “Right” and “Wrong” and it was considered as the response variable.

Full model was fitted using glm() function. Important variables were selected using the backward elimination method. stepAIC() function in R was used to select the variables to the best model. It starts with the full model and backward elimination was done by specifying direction argument as “backward”. It removes variables with highest AIC values and give the final model. Selection procedure was automatically performed by R software. After finding the best model, Hosmer-Lemeshow test was done to find goodness of fit using training data set.

3. Data Exploration

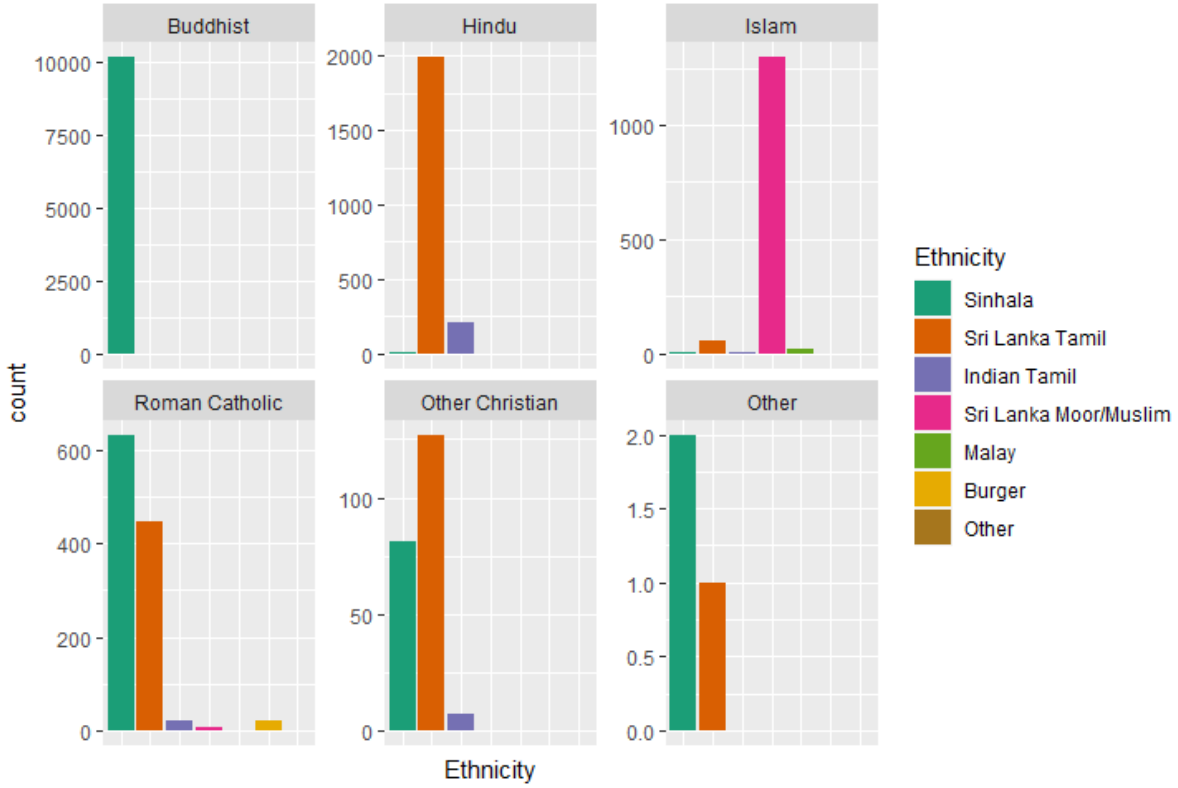
3.1 Composition of the sample

Figure 1: Distribution of ever married women by region and residence



According to figure 1 in each province highest number of women are living in rural areas. Most of the women living in urban areas are in the Western province. In Central and Uva provinces more women are living in estate areas than urban areas. There are no estate areas in Northern, Eastern and North Central provinces.

Figure 2: Distribution of ever married women by religion and ethnicity



All Buddhist women are Sinhalese. Most of the Hindu women are Sri Lanka Tamil. Ethnicity of Islam women are Sri Lanka Moor/Muslim. Among Roman Catholic women majority are Sinhala while second highest ethnicity is Sri Lanka Tamil. Most of other Christians were Sri Lanka Tamil.

Figure 3: Distribution of ever married women by age group

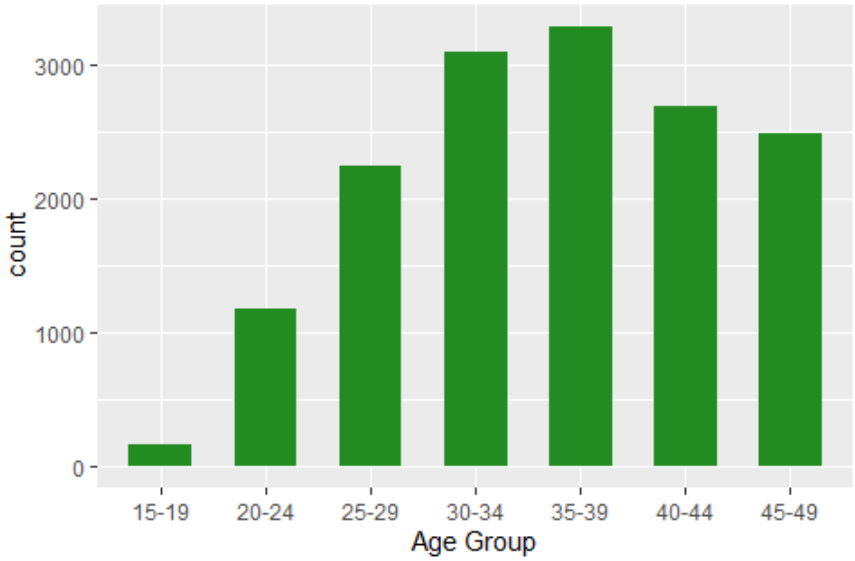
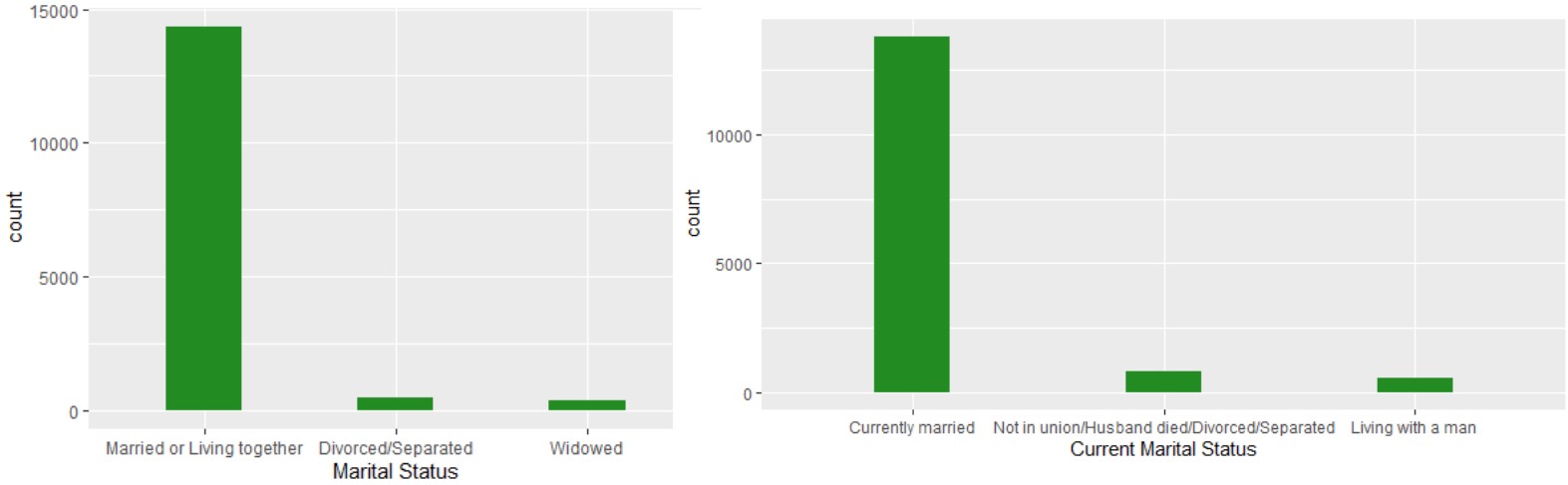


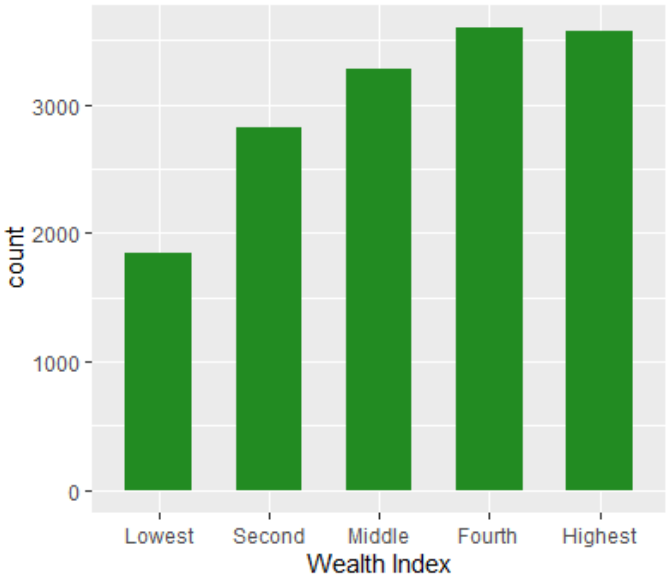
Figure 3 shows that majority of ever married women are aged between 30 to 39 years. Least number of women belong to the age group 15 to 19 years.

Figure 4: Distribution of ever married women by marital status and current marital status



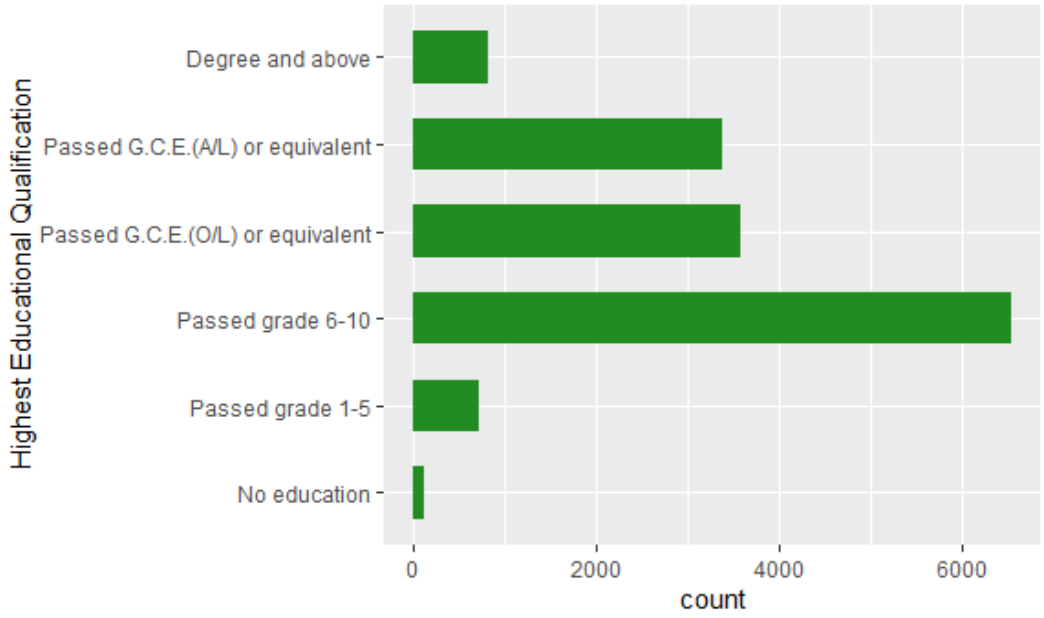
Majority of the women were currently married, and fewer number of women were widowed/divorced/separated or living with a man.

Figure 5: Distribution of ever married women by wealth index



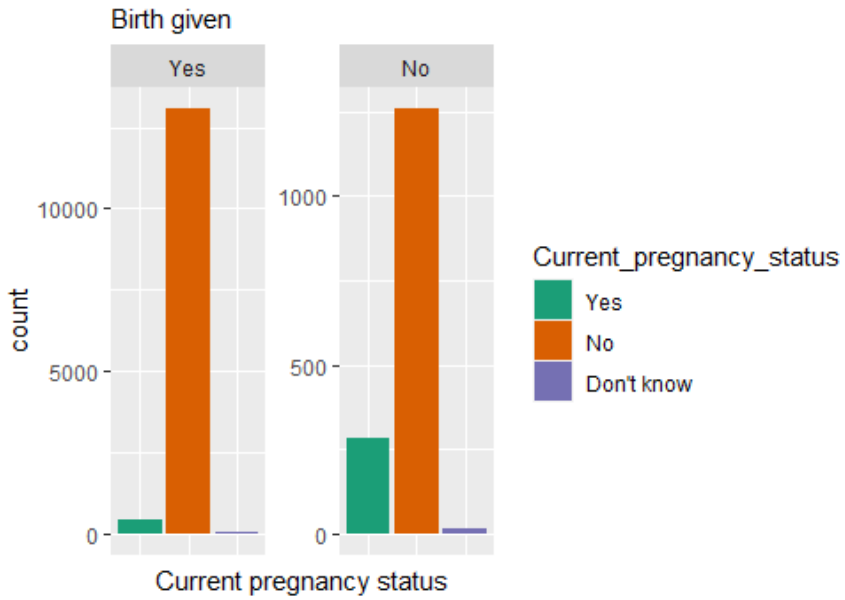
Larger number of women were belonged to fourth and highest wealth index while less than 2000 women were belonged to lowest wealth index.

Figure 6: Distribution of ever married women by highest education qualification



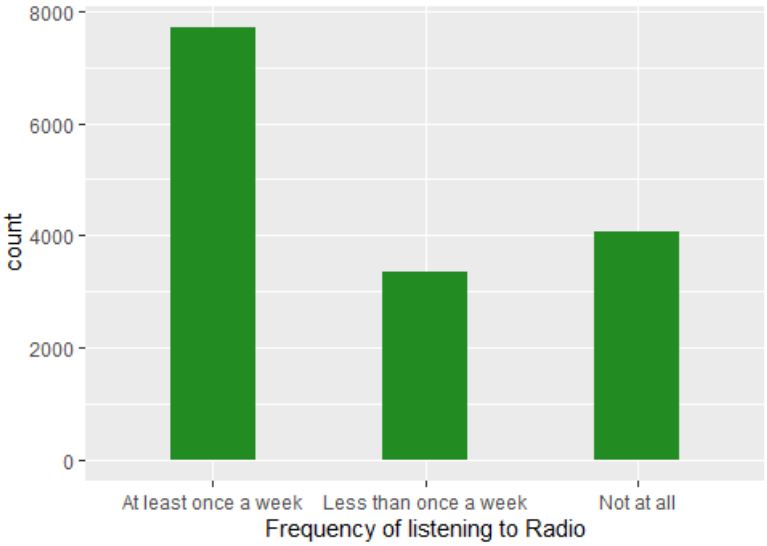
Most women had passed grade 6-10 while almost same number of women had passed G.C.E.(O/L) and G.C.E.(A/L). Very few women had passed grade 1-5 or has a degree.

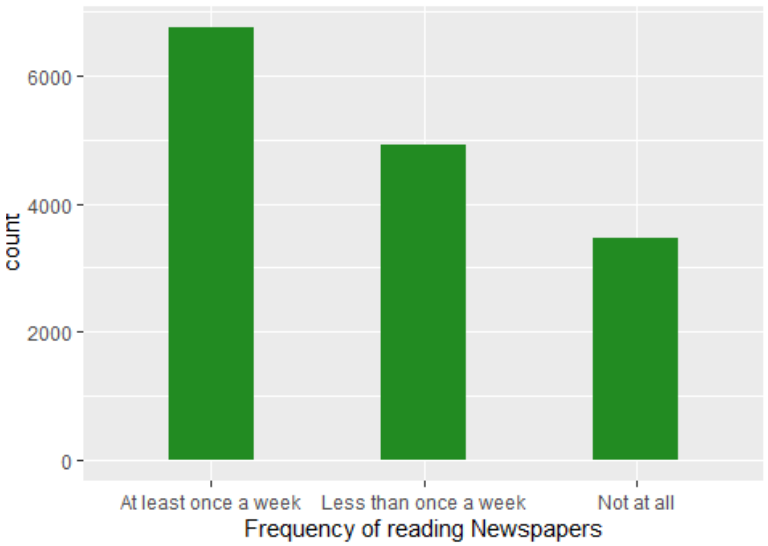
Figure 7: Distribution of ever married women by birth given and current pregnancy status



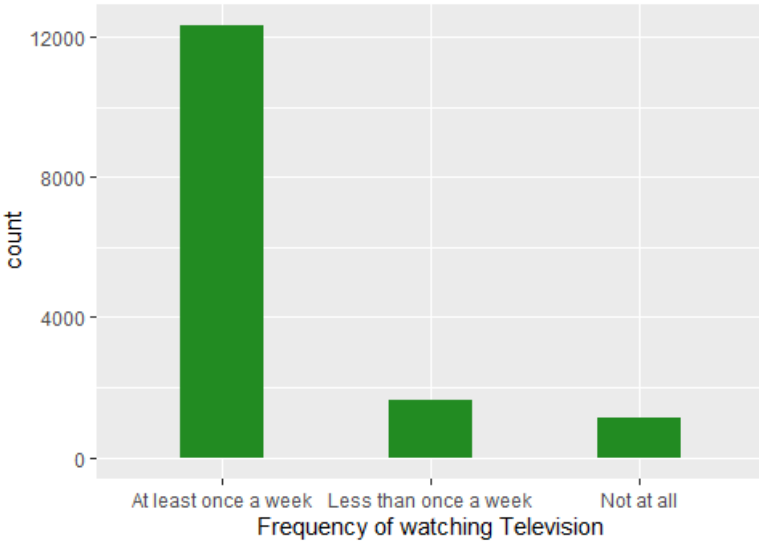
Women who were ever given birth are currently not pregnant. Some women who were not given birth before are currently pregnant.

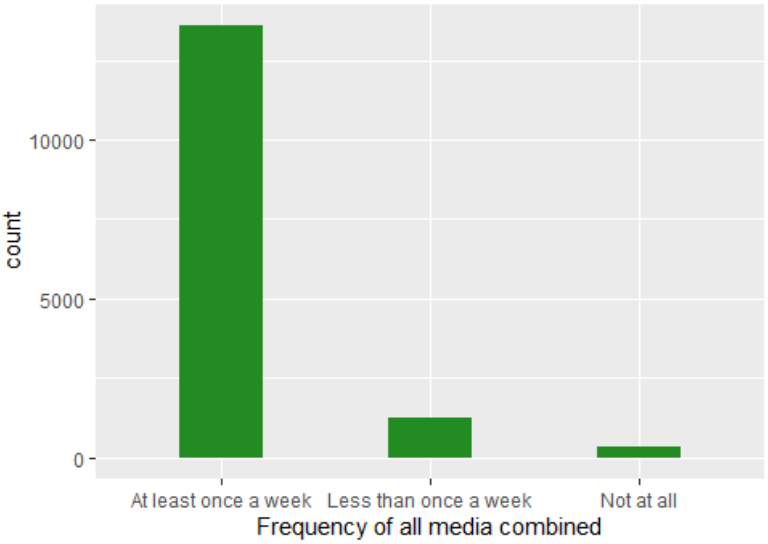
Figure 8: Distribution of ever married women by frequency of radio listening, frequency of





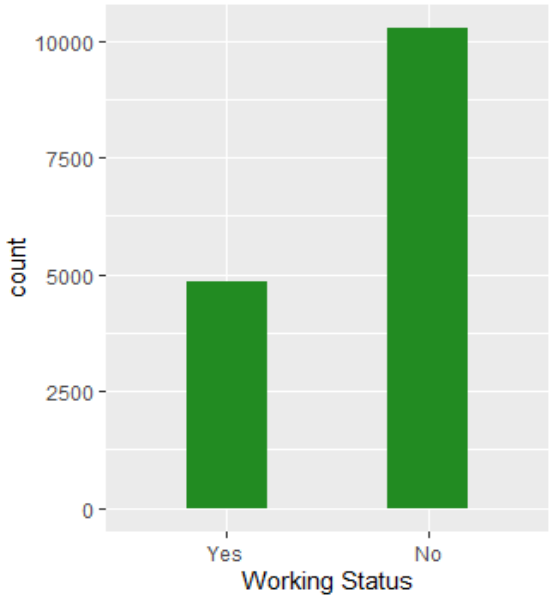
reading newspapers, frequency of watching television and frequency of all media combined





Majority of women were listening to radio, reading newspapers, and watching television at least once a week. There are women who does not listen to radio than women who are listen to radio less than once a week. More women read newspapers less than once a week than not reading newspapers.

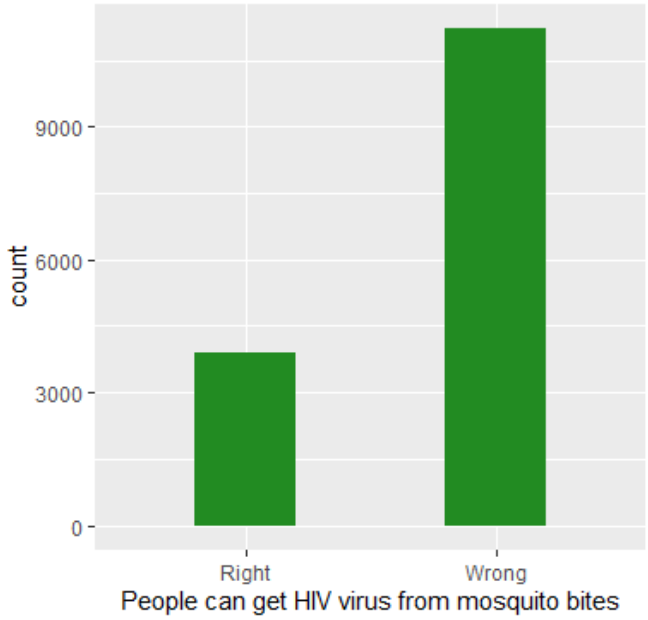
Figure 9: Distribution of ever married women by working status



More than half of the women were not working.

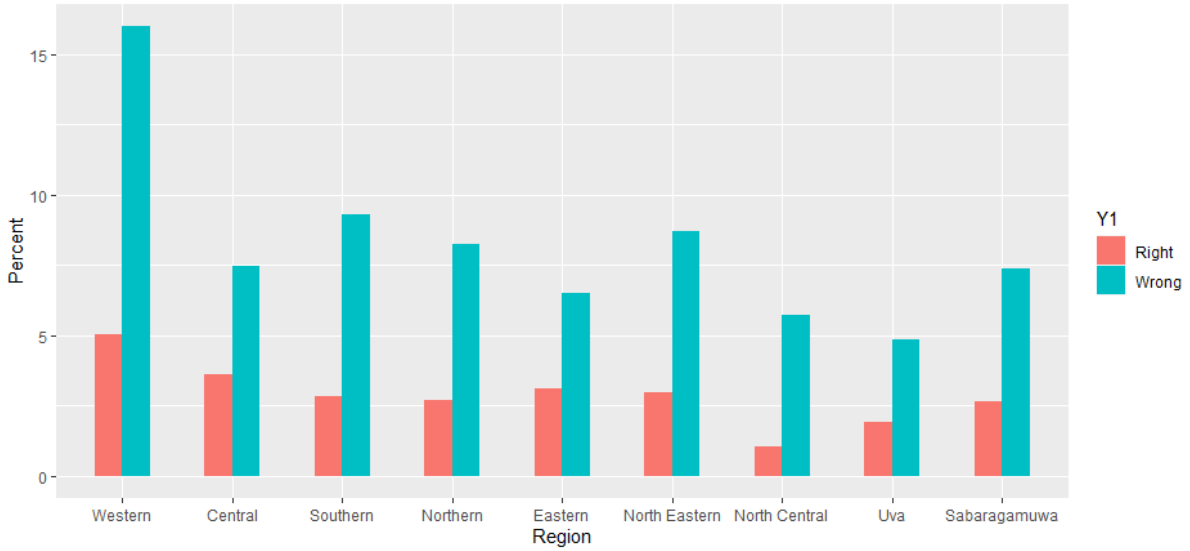
3.2 Analysis of Y1 (response)

Figure 10: Distribution of the response (Y1)



Considerable number of ever married women had responded “wrong” to the myth regarding people getting HIV virus from mosquito bites.

Figure 11: Distribution of Y1 by Region



Majority of women from all the districts had responded wrong to the myth.

Figure 11: Distribution of Y1 by Ethnicity

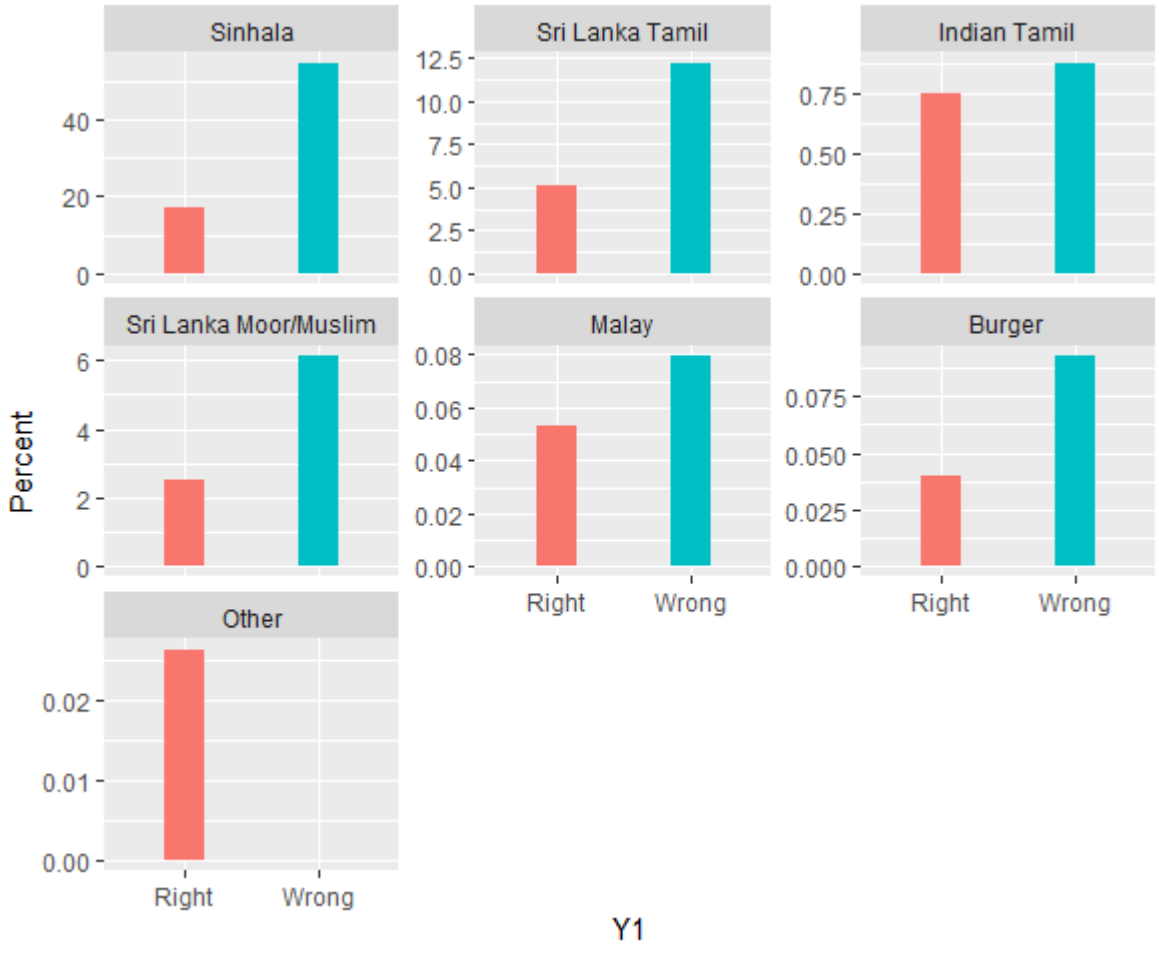
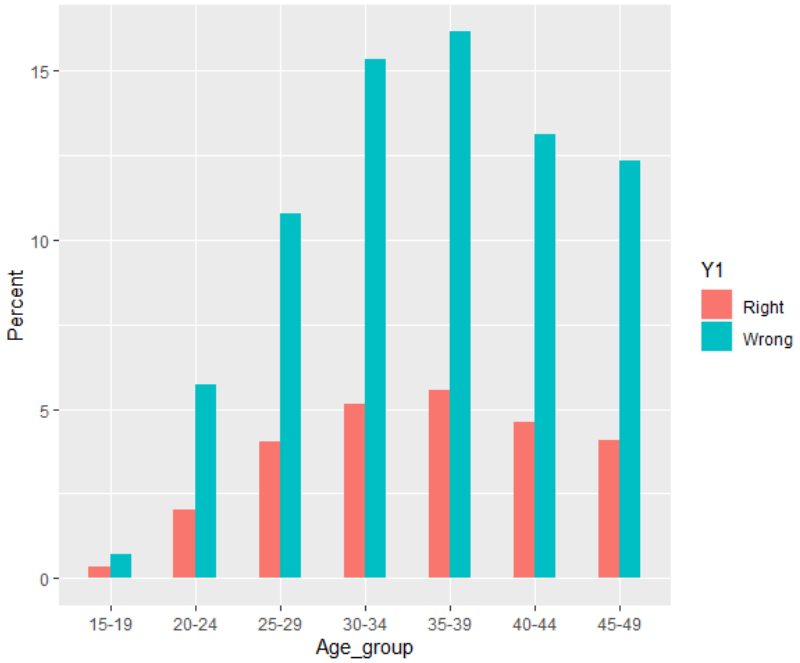


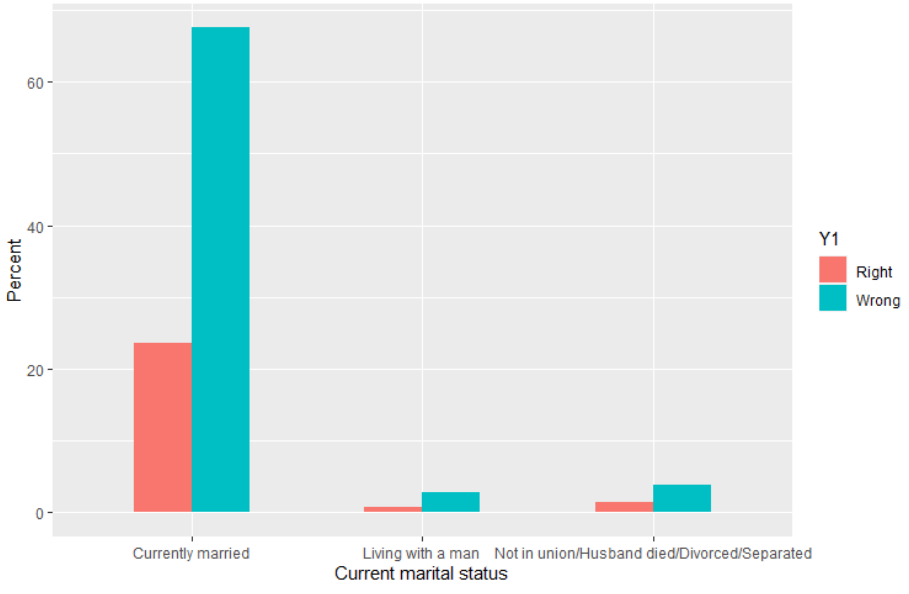
Figure 11 shows that most of the women in every ethnicity had answered wrong to the myth. Number of Indian Tamil women who have answered right to the myth are just below the women who had answered wrong. Women belongs to other ethnicity type had only responded right to the myth.

Figure 12: Distribution of Y1 by Age group



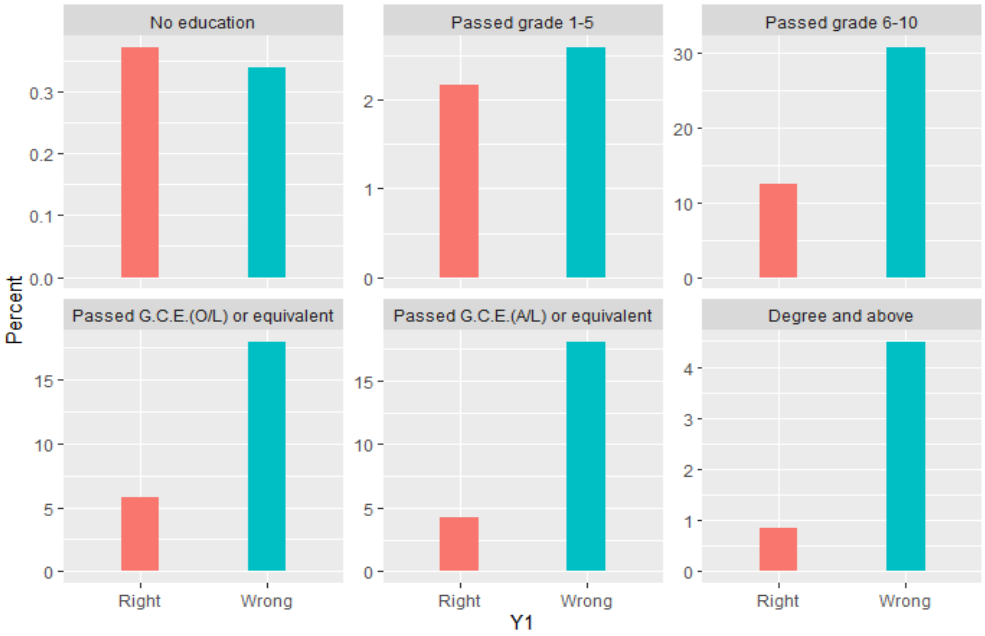
Most women in all age groups had said that the myth is wrong.

Figure 13: Distribution of Y1 by Current marital status



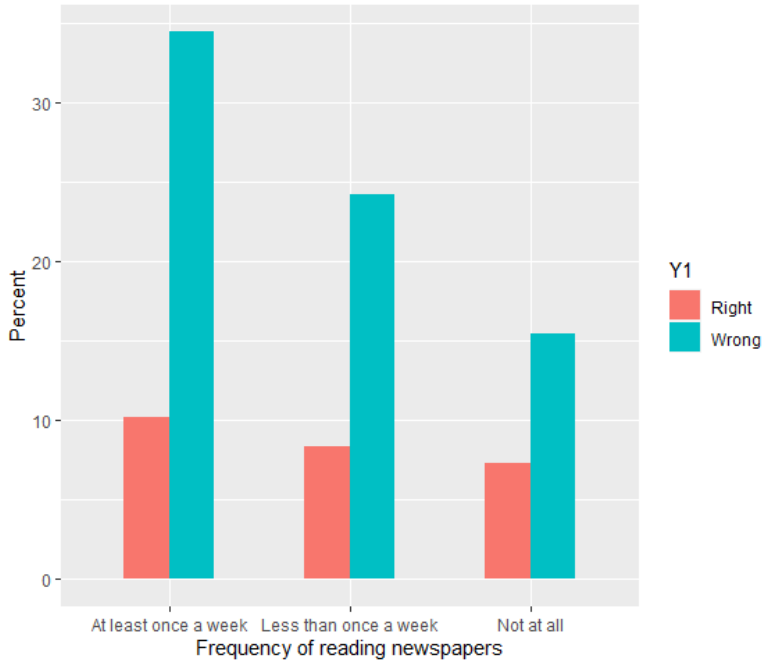
Considerable number of women who are currently married had agreed that the myth is wrong.

Figure 14: Distribution of Y1 by Highest educational qualification



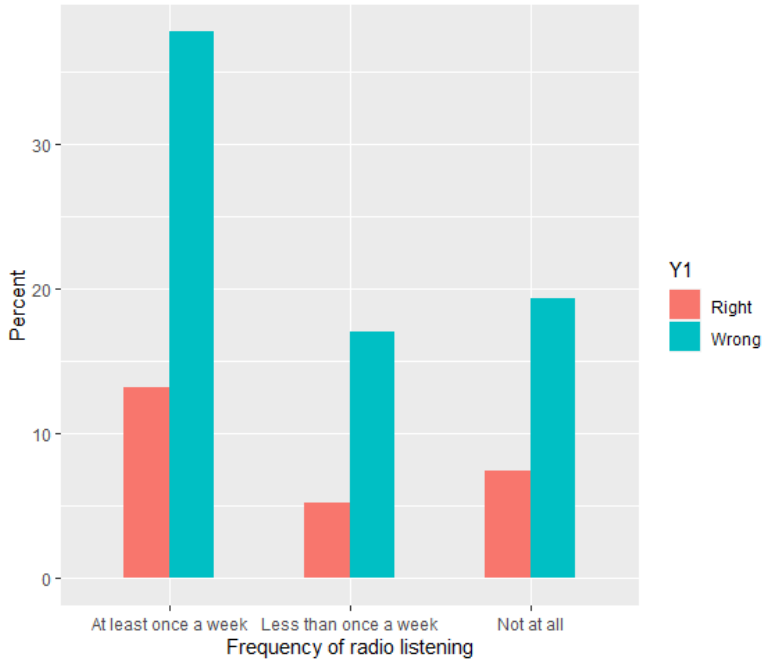
Majority of women who did not had education has answered right to the myth while most of women who had education had answered wrong to the myth.

Figure 14: Distribution of Y1 by Frequency of reading newspapers



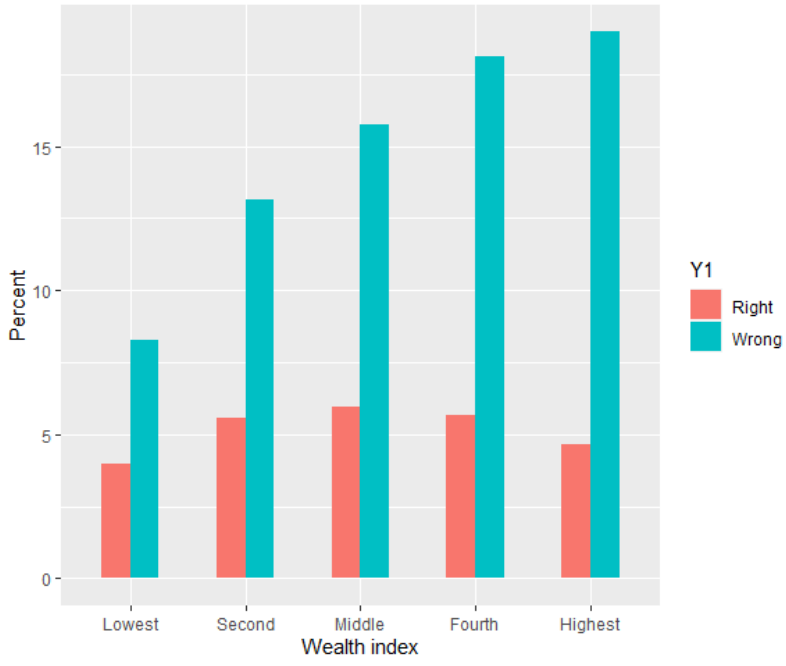
Majority of women who reads newspapers and who does not read newspapers had responded wrong to the myth.

Figure 15: Distribution of Y1 by Frequency of radio listening



Highest number of women who listen to radio and who do not listen to radio has agreed that the myth is wrong.

Figure 16: Distribution of Y1 by Wealth index



Many women in each wealth index had answered wrong to the myth.

4. Data Analysis

4.1 Summary of the data set

```
── Data Summary ────────────────────────
                           Values
Name                       hiv_data
Number of rows             15121
Number of columns          18
_____________________
Column type frequency:
  factor                   17
  numeric                  1
_____________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────────────────────────────────────────────────────────────
─────────────────────────────────
   skim_variable                     n_missing complete_rate ordered n_unique top_counts
 1 Residence                                 0             1 FALSE          3 2: 12046, 1: 2551, 3: 524
 2 Region                                    0             1 FALSE          9 1: 3181, 3: 1836, 6: 1764, 2: 1677
 3 Religion                                  0             1 FALSE          6 1: 10177, 2: 2213, 3: 1382, 4: 1131
 4 Ethnicity                                 0             1 FALSE          7 1: 10894, 2: 2628, 4: 1308, 3: 247
 5 Age_group                                 0             1 FALSE          7 5: 3284, 4: 3096, 6: 2683, 7: 2486
 6 Marital_status                            0             1 FALSE          3 1: 14329, 2: 446, 3: 346
 7 Current_marital_status                    0             1 FALSE          3 1: 13785, 3: 809, 2: 527
 8 Highest_educational_qualification         0             1 FALSE          6 3: 6532, 4: 3581, 5: 3372, 6: 808
 9 Frequency_of_reading_newspapers           0             1 FALSE          3 1: 6749, 2: 4924, 3: 3448
10 Frequency_of_watching_television          0             1 FALSE          3 1: 12324, 2: 1649, 3: 1148
11 Frequency_of_radio_listening              0             1 FALSE          3 1: 7708, 3: 4054, 2: 3359
12 Frequency_of_all_media_combined           0             1 FALSE          3 1: 13579, 2: 1231, 3: 311
13 Given_birth                               0             1 FALSE          2 1: 13569, 2: 1552
14 Current_pregnancy_status                  0             1 FALSE          3 2: 14348, 1: 732, 8: 41
15 Working_status                            0             1 FALSE          2 2: 10268, 1: 4853
16 Wealth_index                              0             1 FALSE          5 4: 3595, 5: 3572, 3: 3281, 2: 2825
17 Y1                                        0             1 FALSE          2 2: 11215, 1: 3906

── Variable type: numeric ─────────────────────────────────────────────────────────────────────────────────────────────────────────────
─────────────────────────────────
  skim_variable n_missing complete_rate  mean    sd p0  p25  p50   p75  p100 hist
1 Woman ID              0             1 9108. 5367.  1 4343 9031 13753 18302 ▇▇▇▇▇▇
```

After removing the missing values and "Don't know" cases of Y1, data frame contains 18 columns and 15121 rows. There are 17 factor variables and one numeric variable. "n_unique" represents the number of levels for each variable. R consider the first level of each variable as the reference level.

Reference levels of each variable are as follows:

No	Variable	Reference Level
1	Residence	Urban
2	Region	Western
3	Religion	Buddhist
4	Ethnicity	Sinhala
5	Age_group	15-19 age group
6	Marital_status	Married or Living together
7	Current_marital_status	Currently married
8	Highest_educational_qualification	No education
9	Frequency_of_reading_newspapers	At least once a week
10	Frequency_of_watching_television	At least once a week
11	Frequency_of_radio_listening	At least once a week

12	Frequency_of_all_media_combined	At least once a week
13	Given_birth	Yes
14	Current_pregnancy_status	Yes
15	Working_status	Yes
16	Wealth_index	Yes
17	Y1	Right

4.2 Variable Selection

Full model was fitted using the training data set. Variables to the best model was selected according to the backward elimination procedure using the full model.

```
> backward$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Y1 ~ Residence + Region + Religion + Ethnicity + Age_group +
    Marital_status + Current_marital_status + Highest_educational_qualification +
    Frequency_of_reading_newspapers + Frequency_of_watching_television +
    Frequency_of_radio_listening + Frequency_of_all_media_combined +
    Given_birth + Current_pregnancy_status + Working_status +
    Wealth_index

Final Model:
Y1 ~ Region + Ethnicity + Age_group + Current_marital_status +
    Highest_educational_qualification + Frequency_of_reading_newspapers +
    Frequency_of_radio_listening + Wealth_index


                                Step Df   Deviance Resid. Df Resid. Dev      AIC
1                                                      12044   13300.92 13406.92
2                          - Religion  5 4.25512513     12049   13305.17 13401.17
3                         - Residence  2 0.02893729     12051   13305.20 13397.20
4 - Frequency_of_watching_television  2 0.35124078     12053   13305.55 13393.55
5         - Current_pregnancy_status  2 1.13971688     12055   13306.69 13390.69
6                       - Given_birth  1 0.21950010     12056   13306.91 13388.91
7                    - Marital_status  2 2.33219464     12058   13309.25 13387.25
8                    - Working_status  1 0.97943051     12059   13310.23 13386.23
9  - Frequency_of_all_media_combined  2 3.18708700     12061   13313.41 13385.41
```

The backward elimination procedure eliminated variables Religion, Residence, Frequency of watching television, Current pregnant status, given birth, Marital status, Working status and Frequency of all media combined which has the highest AIC values.

The best model was fitted using the selected variables Region, Ethnicity, Age group, Current marital status, Highest educational qualification, frequency of reading newspapers, Frequency of radio listening and Wealth index.

```
> summary(fitted_model)

Call:
glm(formula = Y1 ~ Region + Ethnicity + Age_group + Current_marital_status +
    Highest_educational_qualification + Frequency_of_reading_newspapers +
    Frequency_of_radio_listening + Wealth_index, family = binomial,
    data = train_hiv_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2490  -1.1122   0.6792   0.7910   1.5948

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -0.36992    0.30303  -1.221 0.222190
Region2                             -0.21921    0.07886  -2.780 0.005442 **
Region3                              0.11225    0.07958   1.411 0.158377
Region4                              0.52473    0.11101   4.727 2.28e-06 ***
Region5                              0.03280    0.09276   0.354 0.723647
Region6                              0.12699    0.08049   1.578 0.114607
Region7                              0.72218    0.11029   6.548 5.82e-11 ***
Region8                              0.02045    0.09690   0.211 0.832881
Region9                              0.06808    0.08534   0.798 0.425007
Ethnicity2                          -0.36155    0.08510  -4.249 2.15e-05 ***
Ethnicity3                          -0.63121    0.15511  -4.069 4.71e-05 ***
Ethnicity4                          -0.14657    0.08163  -1.795 0.072584 .
Ethnicity5                          -0.65372    0.56325  -1.161 0.245795
Ethnicity6                          -0.69589    0.52051  -1.337 0.181238
Ethnicity7                         -13.86372  160.17037  -0.087 0.931024
Age_group2                           0.38410    0.19464   1.973 0.048446 *
Age_group3                           0.30565    0.18792   1.627 0.103838
Age_group4                           0.40709    0.18614   2.187 0.028739 *
Age_group5                           0.41081    0.18575   2.212 0.026991 *
Age_group6                           0.46817    0.18730   2.500 0.012433 *
Age_group7                           0.53086    0.18834   2.819 0.004824 **
Current_marital_status2              0.32696    0.12775   2.559 0.010484 *
Current_marital_status3              0.08782    0.09615   0.913 0.361044
Highest_educational_qualification2   0.16744    0.23904   0.700 0.483622
Highest_educational_qualification3   0.87094    0.23008   3.785 0.000153 ***
Highest_educational_qualification4   1.01763    0.23422   4.345 1.39e-05 ***
Highest_educational_qualification5   1.26574    0.23734   5.333 9.66e-08 ***
Highest_educational_qualification6   1.47253    0.25899   5.686 1.30e-08 ***
Frequency_of_reading_newspapers2    -0.11686    0.05277  -2.214 0.026807 *
Frequency_of_reading_newspapers3    -0.19546    0.06052  -3.229 0.001240 **
Frequency_of_radio_listening2        0.15765    0.05696   2.768 0.005642 **
Frequency_of_radio_listening3        0.07011    0.05221   1.343 0.179352
Wealth_index2                       -0.04100    0.07757  -0.529 0.597097
Wealth_index3                       -0.02950    0.07992  -0.369 0.712013
Wealth_index4                        0.06082    0.08239   0.738 0.460426
Wealth_index5                        0.21401    0.09136   2.343 0.019153 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13799  on 12096  degrees of freedom
Residual deviance: 13313  on 12061  degrees of freedom
AIC: 13385

Number of Fisher Scoring iterations: 11
```

Best Model:

$$\pi_i = Probability\ that\ the\ i^{th}\ woman\ answered\ \text{"wrong"}\ to\ the\ myth$$

$$\begin{aligned} logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \\ = & -0.37 - 0.23Region_{Central} + 0.11Region_{Southern} \\ & + 0.52Region_{Northern} + 0.03Region_{Eastern} + 0.13Region_{North\ Eastern} \\ & + 0.72Region_{North\ Central} + 0.02Region_{Uva} + 0.06Region_{Sabaragamuwa} \\ & - 0.36Ethnicity_{Sri\ Lanka\ Tamil} - 0.63Ethnicity_{Indian\ Tamil} \\ & - 0.15Ethnicity_{Sri\ Lanka\ Muslim} - 0.65Ethnicity_{Malay} \\ & - 0.69Ethnicity_{Burger} - 13.86Ethnicity_{Other} + 0.38Age_{group\ 20-24} \\ & + 0.31Age_{group\ 25-29} + 0.41Age_{group\ 30-34} + 0.41Age_{group\ 35-39} \\ & + 0.47Age_{group\ 40-44} + 0.53Age_{group\ 45-49} \\ & + 0.33Current\ marital\ status_{living\ with\ a\ man} \\ & + 0.09Current\ marital\ status_{divorced} \\ & + 0.17Highest\ educational\ qualification_{passed\ grade\ 1-5} \\ & + 0.87Highest\ educational\ qualification_{passed\ grade\ 6-10} \\ & + 1.02\ Highest\ educational\ qualification_{passed\ OL} \\ & + 1.27Highest\ educational\ qualification_{passed\ AL} \\ & + 1.47Highest\ educational\ qualification_{degree} \\ & - 0.12Frequency\ of\ reading\ newspapers_{less\ than\ once\ a\ week} \\ & - 0.195Frequency\ of\ reading\ newspapers_{not\ at\ all} \\ & + 0.16Frequency\ of\ radio\ listening_{less\ than\ once\ a\ week} \\ & + 0.07Frequency\ of\ radio\ listening_{not\ at\ all} - 0.04Wealth\ index_{second} \\ & - 0.02Wealth\ index_{middle} + 0.06Wealth\ index_{fourth} \\ & + 0.21Wealth\ index_{highest} \end{aligned}$$

4.3 Goodness of fit of the model

Goodness of fit of the model was checked using the testing data set by applying Hosmer-Lemeshow goodness of fit statistic.

Hypothesis:

$$H_0: Model\ fits\ the\ data\ well \quad vs. \quad H_1: Model\ does\ not\ fit\ the\ data\ well$$

```
> hoslem.test(test_hiv_data$Y1, fitted(model))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  test_hiv_data$Y1, fitted(model)
X-squared = 6.6467, df = 8, p-value = 0.5752
```

P-value is greater than the 5% significance level. Hence the model fits the data well at 5% significance level.

5. Results and Conclusions

The sample consisted of women from all nine provinces and most of them are from rural areas. Majority of women were Sinhalese and Buddhist. Most of the women were aged between 25 to 49 and were currently married. More than 3000 women were belonged to the middle, fourth and highest wealth index. Greater number of ever married women had passed grade 6-10. Majority of the sample is not currently pregnant and do not work. Women were used to listen to radio, read newspapers and watch television at least once a week. Larger number of women had answered “wrong” to the myth regarding HIV. However, all women who belong to other ethnicity had answered “right’’ to the myth. Moreover, majority of women who did not have any education had responded that the myth is right.

Using the fitted logistic regression model, it was found that variables Region, Ethnicity, Age group, Current marital status, Highest educational qualification, frequency of reading newspapers, Frequency of radio listening and Wealth index were the important socio demographic variables that can be used to assess the knowledge regarding the myth. Other variables have been removed due to the high AIC values. Hosmer-Lemeshow test indicated that there is no significant difference between observed and predicted values. Hence fitted model was adequate.

6. References

Cfs.hivci.org. n.d. *HIV Country Profiles*. [online] Available at: https://cfs.hivci.org/

Cichocki, M., 2020. *Can You Get HIV From a Mosquito Bite?*. [online] Verywell Health. Available at: https://www.verywellhealth.com/can-i-get-infected-with-hiv-from-mosquitoes-49547

Who.int. n.d. *HIV/AIDS*. [online] Available at: https://www.who.int/health-topics/hiv-aids#tab=tab_1

Zhang, Z., 2016. Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine*, 4(7), pp.136-136.