# Determinants of Smoking Behavior Among Adults in Bangladesh: A Generalized Linear Model Analysis of Socio-Demographic and Environmental Factors

Rashik Mahmud Orchi

10/04/2024

## Abstract

This study examines the determinants of smoking behavior among adults in Bangladesh using a Generalized Linear Model (GLM). Analyzing socio-demographic and environmental factors, we found gender, education level, and geographical division to be significant predictors of smoking. Specifically, being of the second gender category and higher education levels were associated with increased log odds of smoking. Regional disparities were also evident, with divisions like Dhaka and Mymensingh showing significant differences in smoking behavior. Interestingly, wealth index and other variables such as residence, transport, and healthcare access did not significantly influence smoking odds. The model's AIC value of 479.58 and the well-fitted Dharma residual plots indicate a reliable model. Our findings underscore the need for targeted interventions to address the identified influential factors, contributing to the global effort to reduce tobacco use and its health implications.

## Keywords

Wealth Index, Generalized Linear Model (GLM), Binning, Down-sampling, Logistic Regression, Akaike Information Criterion (AIC), Time Series, Generalized Linear Mixed Model (GLMM)

## Introduction

Bangladesh, with a burgeoning population of 171.2 million (Tobacco Control Research Group, 2023), grapples with a pervasive tobacco epidemic. The prevalence of tobacco use stands at a staggering 43.7% (Tobacco Control Research Group, 2023), reflecting a deeply entrenched public health challenge. In 2018, the repercussions of this widespread tobacco consumption were starkly evident, as approximately 1.5 million adults were afflicted with diseases attributable to tobacco use (Nargis et al., 2022). The impact on the younger demographic is equally alarming, with 61,000 children suffering from ailments linked to secondhand smoke exposure in the same year (Nargis et al., 2022). The mortality toll from tobacco use is profound, accounting for 125.72 thousand deaths in 2018 alone, which represents 13.5% of all-cause mortality (Nargis et al., 2022). This not only signifies a loss of life but also a substantial economic burden. The total economic cost, encompassing both direct healthcare expenses and indirect costs such as lost productivity, amounted to USD 3.61 billion. This economic toll is equivalent to 1.4% of Bangladesh's GDP, underscoring the significant financial strain on the nation's economy (Nargis et al., 2022).

The government of Bangladesh has recognized the gravity of this issue and has been working towards tobacco control. Notably, the Prime Minister Sheikh Hasina has declared an ambitious goal to render Bangladesh tobacco-free by 2040. This commitment is reflected in the increased compliance with the World Health Organization Framework Convention on Tobacco Control (WHO FCTC), the establishment of a dedicated tobacco control cell, and the introduction of a health surcharge on all tobacco products. Despite these efforts, challenges persist, particularly in the form of industry interference in public policy, notably by entities such as British American Tobacco Bangladesh (BATB), in which the state holds a significant share. Addressing these challenges is crucial for the realization of the tobacco-free vision for Bangladesh.

Considering these sobering statistics and the ongoing efforts to combat the tobacco crisis, this paper aims to delve into the multifaceted dimensions of the tobacco problem in Bangladesh, exploring its health, social, and economic impacts, and evaluating the effectiveness of current policies and interventions. To investigate this issue, we chose the dataset, 'Global Adult Tobacco Survey (GATS) – Bangladesh 2017'. This dataset is a nationwide household survey of adults aged 15 or above, using a standard method to collect data on different aspects of tobacco use and exposure. The aim is to understand the patterns, determinants, and consequences of tobacco use in Bangladesh, and the subsequent impact this usage may have on the overall health of the people.

The research questions are:

*1. RQ1 - What are the socio-economic factors associated with tobacco smoking among adults in Bangladesh?*

*2. RQ2 - What is the impact of smoking tobacco in public places on the prevalence of smoking behavior of individuals in Bangladesh?*

Our research question RQ1 seeks to identify and analyze the socio-economic variables that correlate with tobacco smoking among adults in Bangladesh. By understanding these factors, the research can highlight the demographic segments most vulnerable to tobacco use and the socio-economic conditions that facilitate this behavior. This knowledge is crucial for designing targeted interventions and policies to reduce tobacco consumption and its associated health risks. RQ2 on the other hand aims to evaluate the influence of smoking tobacco in public places on individual smoking habits. It explores whether the visibility and social acceptance of smoking in public areas contribute to higher rates of smoking among the population. The findings could inform public health strategies, including the need for stricter enforcement of smoking bans in public places to curb the prevalence of smoking and protect non-smokers from secondhand smoke exposure.

Both questions are integral to developing a comprehensive understanding of the tobacco problem in Bangladesh and formulating effective tobacco control measures. By addressing these questions, our research could provide evidence-based recommendations to policymakers and health practitioners working towards the reduction of tobacco-related harm in the country.

# Data Description

The target population of the survey by GATS Bangladesh 2017 includes all non-institutionalized Bangladeshi men and women ranging from 15 years and older. They were all sampled from the households that they consider their usual place of residence. Here, stratified multi-stage cluster sampling was done through an interview consisting of a household screening component and an individual component administered to the selected respondent. Data was collected using electronic handheld devices. The sample size was 14,880 selected households with 12,783 completed individual interviews.

The questionnaire included core questions about background characteristics, tobacco smoking, electronic cigarettes, smokeless tobacco, cessation, secondhand smoke, economics, media, cigarette pack/picture, and knowledge, attitudes, and perceptions. The dataset contains 516 columns containing unique case ID, interviewer ID, and questions, and 12784 rows with valid responses. For the purpose of our modeling and analysis, we downsampled the data to 77 columns and 451 rows.

# Methodology

This section outlines the research design, population and sample, data collection methods, and data analysis plan employed in this study investigating the relationship between socio-economic factors and smoking behavior in Bangladesh.

## Research Design

This research relies on the quantitative method and is conducted on the basis of a cross sectional research design. Quantitative research obtains numerical data that allows for evaluating the associations between two or more variables. A cross-sectional design is such where data are collected with a subpopulation sample of the population. Through employing this method, we explore the extent of the adoption of smoking and its associations with the socio-economic variables among those Bangladeshis who participated in the 2017 survey.

## Population and Sample

The target group for the following study are all Bangladeshis that are 15 and above in age, encompassing 14,880 households across the country. The dataset of this investigation is Global Adult Tobacco Survey in Bangladesh which took places in year 2017. The survey conducted by the Bangladesh Bureau of Statistics and the National Tobacco Control Cell employs a multistage stratified cluster sampling approach to secure nation representation. Such method includes stratifying the masses into strata (like regions) and then randomly choosing clusters within each stratum. The next step is that there is one or more than one person chosen at random from every nominated household. Participation in the study peaked at 90.8% reaching a final point of 12,783 participants in Bangladesh.

## Data Collection Methods

The data for this study was achieved by way of direct talking (in-person survey) with trained persons who used GATS questionnaire as their instrument. Modules of GATS questionnaire comprises of core and optional sections targeted at different categories of tobacco use submitted by the individuals. These relevant sectioned areas cover the domains of demographics, smoking pattern, knowledge, attitude, medical history and exposure to second hand smoke. Our research examines key modules within the core modules that are relevant to our research problem question, focusing on socio-economic issues and smoking behaviors.

## Data Analysis Plan

This study will employ a two-main approaches to data analysis - 1. Data Cleaning, and 2. Feature Engineering

Data cleaning methods will be implemented to address missing values and inconsistencies within the dataset. Techniques such as binning for continuous variables and down-sampling for variables with a high proportion of missing values are used majorly considering specific data distribution.

Feature engineering is used to build a wealth index, composed of the number of major household possessions and some other variables to calculate the level of social and economic status of a person. In this we can assess the effect of wealth on smoking behavior since we don't have explicit variable to use as reference.

### Statistical Modeling

The main technique of statistical inference used in this study will be logistic regression in order to explore the correlation between social-economic factors (which will include the wealth index that we construct using different variables from dataset as independent variables) and smoking (dependent variable). The use of Generalized Linear Model (GLM), specifically logistic regression is used because of binary nature of our analysis, it is supposed to explore the relationship between smoking status (smokers vs. non-smokers) and indicator variables. This model helps us to find out the strongest links between socio-economic features and smoking.

### Model Validation

The diagnostic tests will be utilized to confirm the model fitting (e.g., residual plots, Q-Q-Plot) and the model's suitability will be assessed. Coefficient values from the GLM are of primary interest, allowing us to assess the significance and direction of associations between independent variables and smoking prevalence.

Various techniques on the selected variables was used to achieve the smallest AIC value for data fitting procedures. The balance between the factor of model complexity and predictive power is the main goal.

### Additional Considerations

As the data is hierarchical (individuals grouped in clusters), we will examine the application of Generalized Linear Mixed Models (GLMMs) to control for random effects of clustering. However, preliminary analyses indicate minimal random effects, leading us to focus on the GLM framework

### Explanation of Statistical Techniques

*Binning:* This technique involves grouping continuous variables into discrete categories based on pre-defined cut-off points. This can be useful for preparing continuous variables for use with statistical models that require categorical data.

*Down-sampling:* This involves removing observations (data points) from a dataset to address situations with imbalanced classes (e.g., a high proportion of missing values). In our case, all the NAN values were removed.

*Logistic Regression:* This is a statistical method used to model the relationship between a binary dependent variable (e.g., smoker vs. non-smoker) and one or more independent variables (e.g., socio-economic factors). It estimates the probability of an event (smoking) occurring based on the values of the independent variables.

*AIC (Akaike Information Criterion):* This is a measure of model complexity that penalizes models for having too many parameters. The goal is to select the model with the lowest AIC value, which balances model fit.

## Literature Review

The use of tobacco by smokers, in particular, remains an area of debate in public health in Bangladesh. The present review inspects the social- economic factors associated with smoking in Bangladesh, and the public-smoking bylaws under this concept. Besides, it comprises the most applicable theoretical constructs, which will act as a guide for the other researchers in this area.

### Socio-Economic Factors and Smoking

A wide range of examinations provides evidence to show that socio-economic determinants and the rate of smoking are correlated strongly. In a study in Bangladesh, they found substantial evidence that is anchored to the economic background of an individual as the cause of smoking habit. Individuals with low income levels and wealth exhibited higher chances of smoking[3]. This evidence is in line with world trends because early childhood poverty is a strong indicator of the likelihood of catching the behavior of smoking and never quitting it early in life.

### Public Smoking Policies and Smoking Behavior

The general introduction of the prohibition of public smoking has become a great tool to fight against smoking. An investigation into the influence of public smoking regulation in Bangladesh resulted in a decline in smoking behaviour after the enactment of such laws[2]. This indicates that limiting smoking in public areas can be a very useful tool for the reduction of overall smoking rates.

### Theoretical Framework

Moreover, smoking behavior presents a range of theories to be able to understand smoking habits. For example, the Health-Belief Model[4] emphasizes how smokers decide on their potential risk of getting the disease or dying young, how serious the disease or death could be, and the level of advantage they think can benefit them by willingly stopping smoking can highly impact the constancy of their behavior. In addition, Social Cognitive Theory[1] highlights the role of social learning and external experiences in adolescents in the emergence and continuation of smoking customs.

```r
rm(list=ls())
```

## 4. Analysis

```r
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
suppressWarnings({
  Bangla_GATS_2017_Public_use_06Spe2018 <- read_excel("Bangla_GATS_2017_Public_use_06Spe2018.xlsx")
})
```

```r
df <- Bangla_GATS_2017_Public_use_06Spe2018 %>%
  select(CASEID,A04,A06A,A06D,A06E,A06G,A06H,A06I,A06J,A06K,A06L,A06M,A06N,A06O,A06P,A06P,A06Q,A06R,A06T
```

```r
summary(df)
```

```
##      CASEID            A04             A06A            A06D     
##  Min.   :100001   Min.   :1.000   Min.   :1.000   Min.   :1.00  
##  1st Qu.:103734   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.00  
##  Median :107541   Median :3.000   Median :1.000   Median :1.00  
##  Mean   :107460   Mean   :3.157   Mean   :1.094   Mean   :1.05  
##  3rd Qu.:111138   3rd Qu.:4.000   3rd Qu.:1.000   3rd Qu.:1.00  
##  Max.   :114880   Max.   :8.000   Max.   :7.000   Max.   :2.00  
##                                                                 
##       A06E            A06G            A06H            A06I      
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000  
##  Median :1.000   Median :2.000   Median :2.000   Median :2.000  
##  Mean   :1.454   Mean   :1.654   Mean   :1.966   Mean   :1.897  
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000  
##  Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :7.000  
##                                                                 
##       A06J            A06K            A06L            A06M           A06N      
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000  
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.00   1st Qu.:1.000  
##  Median :2.000   Median :2.000   Median :2.000   Median :1.00   Median :1.000  
##  Mean   :1.991   Mean   :1.707   Mean   :1.831   Mean   :1.46   Mean   :1.196  
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.00   3rd Qu.:1.000  
##  Max.   :7.000   Max.   :7.000   Max.   :2.000   Max.   :7.00   Max.   :7.000  
##                                                                                
##       A06O            A06P            A06Q            A06R      
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000  
##  Median :1.000   Median :1.000   Median :2.000   Median :2.000  
##  Mean   :1.226   Mean   :1.111   Mean   :1.527   Mean   :1.909  
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:2.000  
##  Max.   :2.000   Max.   :7.000   Max.   :2.000   Max.   :7.000  
##                                                                 
##       A06T            A06U            A14             A06B     
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00  
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00  
##  Median :2.000   Median :2.000   Median :2.000   Median :2.00  
##  Mean   :1.665   Mean   :1.941   Mean   :2.147   Mean   :1.91  
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.00  
##  Max.   :2.000   Max.   :9.000   Max.   :3.000   Max.   :7.00  
##                                                                
##       HH2             HH4E01          HH4B01           HH4D01     
##  Min.   : 1.000   Min.   :1.000   Min.   : 15.00   Min.   :1.000  
##  1st Qu.: 2.000   1st Qu.:2.000   1st Qu.: 33.00   1st Qu.:1.000  
##  Median : 3.000   Median :2.000   Median : 42.00   Median :2.000  
##  Mean   : 3.029   Mean   :1.811   Mean   : 44.22   Mean   :1.524  
##  3rd Qu.: 4.000   3rd Qu.:2.000   3rd Qu.: 55.00   3rd Qu.:2.000  
##  Max.   :20.000   Max.   :7.000   Max.   :105.00   Max.   :2.000  
## 
```

```
##    RESIDENCE          E12            E14             E16             E20
##  Min.   :1.000   Min.   :1.0    Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.0    1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :2.000   Median :2.0    Median :1.000   Median :2.000   Median :2.000
##  Mean   :1.503   Mean   :2.1    Mean   :1.562   Mean   :1.706   Mean   :2.004
##  3rd Qu.:2.000   3rd Qu.:2.0    3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :9.0    Max.   :7.000   Max.   :7.000   Max.   :7.000
##                  NA's   :8936   NA's   :8730    NA's   :5862    NA's   :9876
##       E11             E13             E15             E19           divisionid
##  Min.   :1.000   Min.   :1.00   Min.   :1.00   Min.   :1.000   Min.   :10.00
##  1st Qu.:1.000   1st Qu.:1.00   1st Qu.:1.00   1st Qu.:2.000   1st Qu.:30.00
##  Median :2.000   Median :2.00   Median :1.00   Median :2.000   Median :45.00
##  Mean   :1.711   Mean   :1.69   Mean   :1.47   Mean   :1.784   Mean   :38.85
##  3rd Qu.:2.000   3rd Qu.:2.00   3rd Qu.:2.00   3rd Qu.:2.000   3rd Qu.:50.00
##  Max.   :9.000   Max.   :9.00   Max.   :7.00   Max.   :9.000   Max.   :60.00
##
##   gatsstrata   gatscluster      gatsweight
##  Min.   :101   Min.   :  1.0   Min.   :   325.2
##  1st Qu.:301   1st Qu.:125.0   1st Qu.:  1984.3
##  Median :451   Median :252.0   Median :  5864.1
##  Mean   :390   Mean   :249.2   Mean   :  8368.8
##  3rd Qu.:502   3rd Qu.:372.0   3rd Qu.: 10738.2
##  Max.   :602   Max.   :496.0   Max.   :207264.2
##
```

```
#######Wealth index
###### New columns have be mutated
#####Previously the value of data was 1& 2
######So it has been converted to binary.


df <- df %>%
  mutate(
    Electricity = ifelse(A06A == 1, 1, 0),
    Flush_toilet = ifelse(A06B == 1, 1, 0),
    Mobile_phone = ifelse(A06D == 1, 1, 0),
    Television = ifelse(A06E == 1, 1, 0),
    refrigerator = ifelse(A06G == 1, 1, 0),
    car_rickshaw= ifelse(A06H == 1, 1, 0),
    cycle_motorcyclce = ifelse(A06I == 1, 1, 0),
    washing_machine = ifelse(A06J == 1, 1, 0),
    Almirah_Wardrobe = ifelse(A06M == 1, 1, 0),
    table = ifelse(A06N == 1, 1, 0),
    bed = ifelse(A06O == 1, 1, 0),
    chair = ifelse(A06P == 1, 1, 0),
    Computer_Laptop_Tab = ifelse(A06R == 1, 1, 0),
    domestic_animal = ifelse(A06T == 1, 1, 0),
    rooftop_material = ifelse(A14 %in% c(1, 2), 0, 1)
  )
```

```
#####PCA###########
library(psych)

asset_variables <- select(df, c(
```

```r
  Electricity, Flush_toilet, Mobile_phone, Television, refrigerator,
  car_rickshaw, cycle_motorcyclce, washing_machine, Almirah_Wardrobe,
  table, bed, chair, Computer_Laptop_Tab, domestic_animal, rooftop_material
))
```

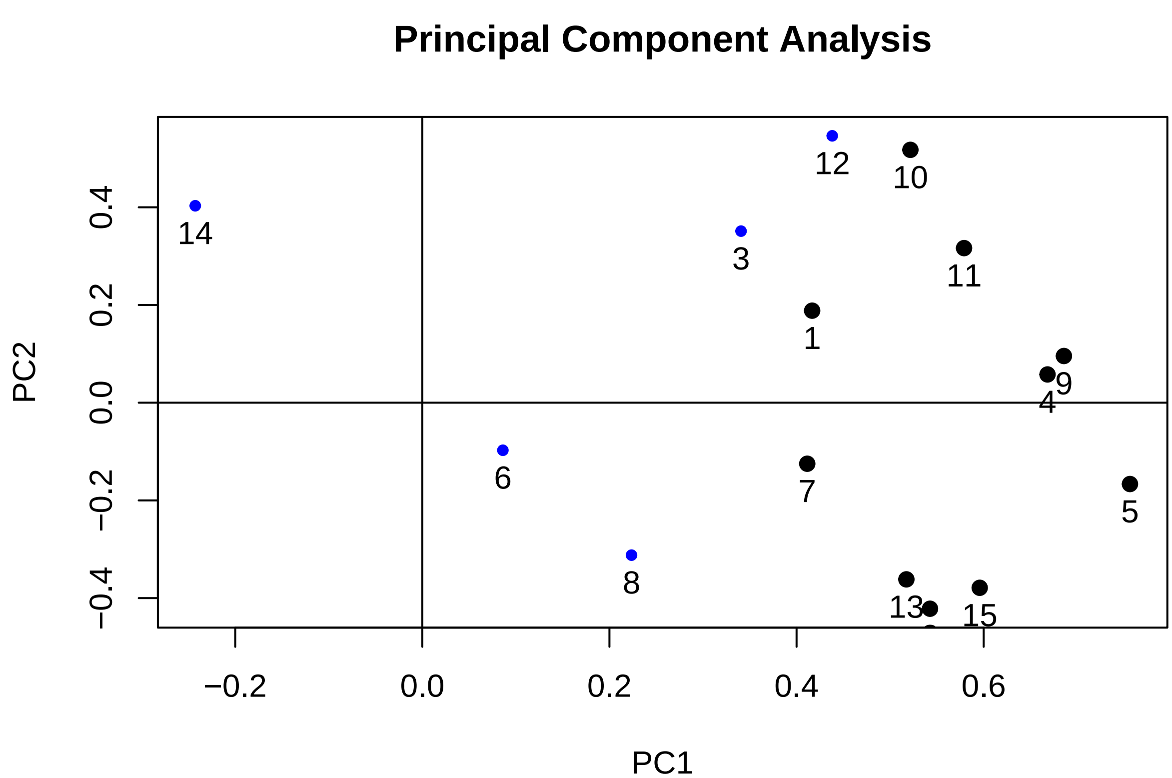
```r
# Since the variables are binary therefore we calculated through MCA

# Performing MCA using the principal() function
mca_result <- principal(asset_variables, nfactors = 2, rotate = "none")

 # Summary of the MCA results
summary(mca_result)
```

```
##
## Factor analysis with Call: principal(r = asset_variables, nfactors = 2, rotate = "none")
##
## Test of the hypothesis that 2 factors are sufficient.
## The degrees of freedom for the model is 76  and the objective function was  0.45
## The number of observations was  12783  with Chi Square =  5714.4  with prob <  0
##
## The root mean square of the residuals (RMSA) is  0.07
```

```r
plot(mca_result)
```



```r
df <- df %>%
  mutate(
    wealth_index = as.vector(mca_result$scores[, 1])
  )
```

```r
##Creating hospital,restaurant, transport,school column
## The vales have been converted to 0 and 1.
df <- df %>%
  mutate(
    hospital = case_when(
      E11 == 1 & E12 == 2 ~ 0,
      E11 == 1 & E12 == 1 ~ 1,
      TRUE ~ NA_real_
    ),
    restaurant = case_when(
      E13 == 1 & E14 == 2 ~ 0,
      E13 == 1 & E14 == 1 ~ 1,
      TRUE ~ NA_real_
    ),
    transport = case_when(
      E15 == 1 & E16 == 2 ~ 0,
      E15 == 1 & E16 == 1 ~ 1,
      TRUE ~ NA_real_
    ),
    school = case_when(
      E19 == 1 & E20 == 2 ~ 0,
      E19 == 1 & E20 == 1 ~ 1,
      TRUE ~ NA_real_
    )
  )
```

```r
##############AGE group
### We did binning of the continous variable HH4B01(age)
breaks <- c(15, 35, 50, 65, 80, 105)


labels <- c("15-34", "35-49", "50-64", "65-79", "80-105")


df <- df %>%
  mutate(age_group = cut(HH4B01, breaks = breaks, labels = labels, include.lowest = TRUE))
```

```r
###########education level######
###########education level######
df <- df %>%
  mutate(education = case_when(
    A04 == 1 ~ "No Schooling",
    A04 %in% c(2, 3) ~ "atbest_Primary",
    A04 %in% c(4, 5) ~ "atbest_Secondary",
    A04 %in% c(6, 7, 8) ~ "Higher_Education",
    TRUE ~ NA_character_
  ))
```

```r
#####We have set the order of the education variable
df$education <- factor(df$education, levels = c("No Schooling", "atbest_Primary", "atbest_Secondary", "I
```

```r
############

df <- df %>%
  mutate(Division_Name = case_when(
    divisionid == 10 ~ "Barisal",
    divisionid == 20 ~ "Chittagong",
    divisionid == 30 ~ "Dhaka",
    divisionid == 40 ~ "Khulna",
    divisionid == 45 ~ "Mymensingh",
    divisionid == 50 ~ "Rajshahi",
    divisionid == 55 ~ "Rangpur",
    divisionid == 60 ~ "Sylhet",
    TRUE ~ as.character(divisionid)
  ))
```

```r
############Residence########
names(df)[names(df) == "HH4D01"] <- "gender"
df$gender <- as.factor(df$gender)


df$RESIDENCE <- as.factor(df$RESIDENCE)
```

```r
####Down-sampling Based on our criteria


######smoking

###removing smoking that are invalid
df <- df %>%
  mutate(smoking = ifelse(HH4E01 == 2, 0, 1)) %>%
  filter(HH4E01 != 7)

#######Dropping observations that don't satsify the requirements of our research design
new_df <- df %>%
  filter(!is.na(hospital) & !is.na(school) & !is.na(transport) & !is.na(restaurant))
```

```r
new_df$Division_Name= as.factor(new_df$Division_Name)
```

```r
summary(new_df)
```

```
##      CASEID            A04              A06A             A06D
##  Min.   :100220   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:104676   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :107566   Median :5.000   Median :1.000   Median :1.000
##  Mean   :107810   Mean   :4.792   Mean   :1.053   Mean   :1.011
##  3rd Qu.:111800   3rd Qu.:6.000   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :114840   Max.   :8.000   Max.   :2.000   Max.   :2.000
##
##       A06E             A06G             A06H             A06I
```

```
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000  
##  Median :1.000   Median :1.000   Median :2.000   Median :2.000  
##  Mean   :1.239   Mean   :1.441   Mean   :1.962   Mean   :1.745  
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000  
##  Max.   :2.000   Max.   :7.000   Max.   :2.000   Max.   :2.000  
##                                                                 
##       A06J            A06K          A06L            A06M            A06N      
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000  
##  1st Qu.:2.000   1st Qu.:1.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000  
##  Median :2.000   Median :2.00   Median :2.000   Median :1.000   Median :1.000  
##  Mean   :1.976   Mean   :1.63   Mean   :1.754   Mean   :1.215   Mean   :1.064  
##  3rd Qu.:2.000   3rd Qu.:2.00   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:1.000  
##  Max.   :2.000   Max.   :2.00   Max.   :2.000   Max.   :2.000   Max.   :2.000  
##                                                                                
##       A06O            A06P            A06Q            A06R      
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000  
##  Median :1.000   Median :1.000   Median :1.000   Median :2.000  
##  Mean   :1.089   Mean   :1.033   Mean   :1.286   Mean   :1.783  
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:2.000  
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000  
##                                                                 
##       A06T            A06U            A14             A06B      
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000  
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000  
##  Mean   :1.705   Mean   :1.885   Mean   :2.275   Mean   :1.796  
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000  
##  Max.   :2.000   Max.   :2.000   Max.   :3.000   Max.   :2.000  
##                                                                 
##       HH2             HH4E01          HH4B01       gender  RESIDENCE
##  Min.   : 1.000   Min.   :1.000   Min.   : 15.00   1:344   1:300    
##  1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 34.00   2:107   2:151    
##  Median : 3.000   Median :2.000   Median : 42.00                    
##  Mean   : 3.184   Mean   :1.707   Mean   : 43.24                    
##  3rd Qu.: 4.000   3rd Qu.:2.000   3rd Qu.: 52.00                    
##  Max.   :20.000   Max.   :2.000   Max.   :100.00                    
##                                                                     
##       E12             E14             E16             E20             E11   
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1  
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1  
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000   Median :1  
##  Mean   :1.858   Mean   :1.534   Mean   :1.523   Mean   :1.918   Mean   :1  
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1  
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :1  
##                                                                             
##       E13         E15         E19      divisionid     gatsstrata 
##  Min.   :1   Min.   :1   Min.   :1   Min.   :10.00   Min.   :101  
##  1st Qu.:1   1st Qu.:1   1st Qu.:1   1st Qu.:30.00   1st Qu.:302  
##  Median :1   Median :1   Median :1   Median :45.00   Median :451  
##  Mean   :1   Mean   :1   Mean   :1   Mean   :40.53   Mean   :407  
##  3rd Qu.:1   3rd Qu.:1   3rd Qu.:1   3rd Qu.:55.00   3rd Qu.:551  
##  Max.   :1   Max.   :1   Max.   :1   Max.   :60.00   Max.   :602  
```

```
##
##   gatscluster      gatsweight        Electricity      Flush_toilet
##  Min.   :  8.0   Min.   :   411.4   Min.   :0.0000   Min.   :0.000
##  1st Qu.:156.5   1st Qu.:  1547.5   1st Qu.:1.0000   1st Qu.:0.000
##  Median :253.0   Median :  3925.7   Median :1.0000   Median :0.000
##  Mean   :260.8   Mean   :  7640.1   Mean   :0.9468   Mean   :0.204
##  3rd Qu.:394.0   3rd Qu.:  9156.5   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :495.0   Max.   :204910.4   Max.   :1.0000   Max.   :1.000
##
##   Mobile_phone      Television      refrigerator     car_rickshaw
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :0.00000
##  Mean   :0.9889   Mean   :0.7605   Mean   :0.5698   Mean   :0.03769
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##
##  cycle_motorcyclce washing_machine   Almirah_Wardrobe     table
##  Min.   :0.000     Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000     1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:1.0000
##  Median :0.000     Median :0.00000   Median :1.0000   Median :1.0000
##  Mean   :0.255     Mean   :0.02439   Mean   :0.7849   Mean   :0.9357
##  3rd Qu.:1.000     3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000     Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##
##       bed              chair        Computer_Laptop_Tab domestic_animal
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000      Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000      1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :0.0000      Median :0.0000
##  Mean   :0.9113   Mean   :0.9667   Mean   :0.2173      Mean   :0.2949
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000      3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000      Max.   :1.0000
##
##  rooftop_material  wealth_index         hospital        restaurant
##  Min.   :0.0000   Min.   :-2.09597   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:-0.08973   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median : 0.68551   Median :0.0000   Median :0.0000
##  Mean   :0.2927   Mean   : 0.64461   Mean   :0.1419   Mean   :0.4656
##  3rd Qu.:1.0000   3rd Qu.: 1.37973   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   : 3.04644   Max.   :1.0000   Max.   :1.0000
##
##    transport          school          age_group             education
##  Min.   :0.0000   Min.   :0.00000   15-34 :139   No Schooling    : 34
##  1st Qu.:0.0000   1st Qu.:0.00000   35-49 :193   atbest_Primary  : 83
##  Median :0.0000   Median :0.00000   50-64 : 95   atbest_Secondary:150
##  Mean   :0.4767   Mean   :0.08204   65-79 : 20   Higher_Education:184
##  3rd Qu.:1.0000   3rd Qu.:0.00000   80-105:  4
##  Max.   :1.0000   Max.   :1.00000
##
##     Division_Name    smoking
##  Khulna    :93    Min.   :0.0000
##  Rangpur   :83    1st Qu.:0.0000
##  Mymensingh:55    Median :0.0000
##  Barisal   :48    Mean   :0.2927
```

```
##  Sylhet    :47    3rd Qu.:1.0000
##  Dhaka     :43    Max.   :1.0000
##  (Other)   :82
```

```r
########with age_group binned
m1 <- glm(smoking~age_group+gender+wealth_index+RESIDENCE+transport+school+hospital+restaurant+educatio
summary(m1)
```

```
##
## Call:
## glm(formula = smoking ~ age_group + gender + wealth_index + RESIDENCE +
##     transport + school + hospital + restaurant + education +
##     Division_Name, family = "binomial", data = new_df)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.47733    0.50469  -2.927  0.00342 **
## age_group35-49            0.58020    0.30369   1.910  0.05607 .
## age_group50-64            0.64886    0.33659   1.928  0.05389 .
## age_group65-79           -0.08969    0.59943  -0.150  0.88106
## age_group80-105         -14.03481  675.79239  -0.021  0.98343
## gender2                  -4.46070    1.03888  -4.294 1.76e-05 ***
## wealth_index             -0.08963    0.16576  -0.541  0.58870
## RESIDENCE2                0.25877    0.26783   0.966  0.33396
## transport                 0.18675    0.24595   0.759  0.44768
## school                    0.67291    0.40480   1.662  0.09645 .
## hospital                  0.16717    0.35260   0.474  0.63541
## restaurant               -0.03587    0.25364  -0.141  0.88755
## education.L              -1.12451    0.38882  -2.892  0.00383 **
## education.Q               0.47653    0.32403   1.471  0.14138
## education.C              -0.06488    0.25751  -0.252  0.80108
## Division_NameChittagong   0.19419    0.59270   0.328  0.74319
## Division_NameDhaka        1.10066    0.53827   2.045  0.04087 *
## Division_NameKhulna       0.66913    0.48051   1.393  0.16375
## Division_NameMymensingh   1.03201    0.49229   2.096  0.03605 *
## Division_NameRajshahi     0.82759    0.55630   1.488  0.13684
## Division_NameRangpur      0.90569    0.47997   1.887  0.05916 .
## Division_NameSylhet       0.68531    0.52607   1.303  0.19268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 545.29  on 450  degrees of freedom
## Residual deviance: 435.58  on 429  degrees of freedom
## AIC: 479.58
##
## Number of Fisher Scoring iterations: 14
```

```r
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'

## The following object is masked from 'package:psych':
##
##     logit

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
# Fit the logistic regression model
m1 <- glm(smoking ~ age_group + gender + wealth_index + RESIDENCE + transport + school + hospital + res

# Check for multicollinearity using VIF
vif_results <- vif(m1)
print(vif_results)
```

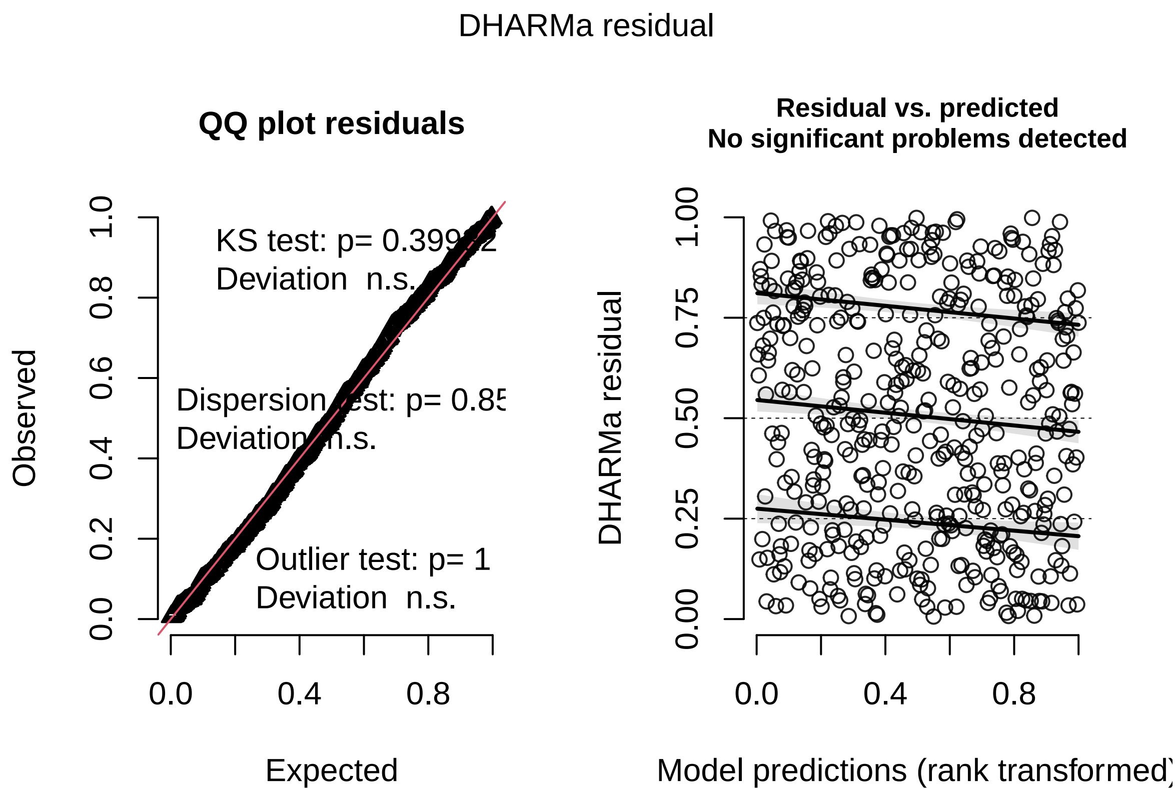
```
##                  GVIF Df GVIF^(1/(2*Df))
## age_group     1.251769  4        1.028467
## gender        1.045914  1        1.022699
## wealth_index  1.868516  1        1.366937
## RESIDENCE     1.208892  1        1.099496
## transport     1.132062  1        1.063984
## school        1.161203  1        1.077591
## hospital      1.157787  1        1.076005
## restaurant    1.193200  1        1.092337
## education     1.704651  3        1.092964
## Division_Name 1.767212  7        1.041510
```

```
library("DHARMa")
```

```
## This is DHARMa 0.4.6. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa')
```

```
residuals <- simulateResiduals(m1)

plot(residuals)
```




```r
# Creating a survey design object to include the survey weight
###install.packages("survey")
library(survey)
```

```
## Loading required package: grid

## Loading required package: Matrix

## Loading required package: survival

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart
```

```r
survey_design <- suppressWarnings(svydesign(ids = ~1, weights = ~gatsweight, data = new_df))
```

```r
# Estimating representation of each observation of the total populaton

total_population <- sum(weights(survey_design))
```

```r
representation <- total_population / nrow(new_df)


print(representation)
```

```
## [1] 7640.112
```

So, each obeservation in the dataset represent 7460 person in real life.

```r
###Model Incorporating Survey Weight

m1_survey <- svyglm(smoking ~ age_group + gender + wealth_index + RESIDENCE + transport + school + hospi
              design = survey_design, family = "binomial")
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```r
# Summarize the model results
summary(m1_survey)
```

```
##
## Call:
## svyglm(formula = smoking ~ age_group + gender + wealth_index +
##     RESIDENCE + transport + school + hospital + restaurant +
##     education + Division_Name, design = survey_design, family = "binomial")
##
## Survey design:
## svydesign(ids = ~1, weights = ~gatsweight, data = new_df)
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2.49020    0.80545  -3.092  0.00212 **
## age_group35-49             1.21891    0.51864   2.350  0.01921 *
## age_group50-64             0.93918    0.50573   1.857  0.06398 .
## age_group65-79             0.57732    0.72648   0.795  0.42724
## age_group80-105          -15.53016    0.99244 -15.648  < 2e-16 ***
## gender2                   -5.42651    1.07992  -5.025 7.41e-07 ***
## wealth_index              -0.05436    0.25922  -0.210  0.83400
## RESIDENCE2                 0.63886    0.40050   1.595  0.11141
## transport                  0.50772    0.37060   1.370  0.17140
## school                     2.12520    0.68624   3.097  0.00208 **
## hospital                   0.25699    0.54449   0.472  0.63718
## restaurant                -0.62128    0.40916  -1.518  0.12965
## education.L               -1.19975    0.50959  -2.354  0.01900 *
## education.Q                1.13568    0.45973   2.470  0.01389 *
## education.C                0.02778    0.41725   0.067  0.94696
## Division_NameChittagong   -0.23318    0.79887  -0.292  0.77052
## Division_NameDhaka         1.11788    0.71852   1.556  0.12049
## Division_NameKhulna        1.44948    0.73627   1.969  0.04963 *
## Division_NameMymensingh    1.62459    0.67579   2.404  0.01664 *
## Division_NameRajshahi      2.13734    0.82121   2.603  0.00957 **
## Division_NameRangpur       0.73176    0.71471   1.024  0.30649
```

```
## Division_NameSylhet       1.27452    1.00042   1.274  0.20336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1.120524)
##
## Number of Fisher Scoring iterations: 15
```

```r
step(m1)
```

```
## Start:  AIC=479.58
## smoking ~ age_group + gender + wealth_index + RESIDENCE + transport +
##     school + hospital + restaurant + education + Division_Name
##
##                 Df Deviance    AIC
## - Division_Name  7   443.60 473.60
## - restaurant     1   435.60 477.60
## - hospital       1   435.81 477.81
## - wealth_index   1   435.88 477.88
## - transport      1   436.16 478.16
## - RESIDENCE      1   436.51 478.51
## <none>               435.58 479.58
## - age_group      4   443.68 479.68
## - school         1   438.34 480.34
## - education      3   444.65 482.65
## - gender         1   506.84 548.84
##
## Step:  AIC=473.6
## smoking ~ age_group + gender + wealth_index + RESIDENCE + transport +
##     school + hospital + restaurant + education
##
##                Df Deviance    AIC
## - hospital      1   443.66 471.66
## - restaurant    1   443.68 471.68
## - wealth_index  1   444.17 472.17
## - transport     1   444.34 472.34
## - RESIDENCE     1   444.84 472.84
## - school        1   445.53 473.53
## <none>              443.60 473.60
## - age_group     4   451.62 473.62
## - education     3   451.50 475.50
## - gender        1   516.10 544.10
##
## Step:  AIC=471.66
## smoking ~ age_group + gender + wealth_index + RESIDENCE + transport +
##     school + restaurant + education
##
##                Df Deviance    AIC
## - restaurant    1   443.73 469.73
## - wealth_index  1   444.19 470.19
## - transport     1   444.46 470.46
## - RESIDENCE     1   444.98 470.98
## <none>              443.66 471.66
## - school        1   445.72 471.72
```

```
## - age_group     4   451.79 471.79
## - education     3   451.55 473.55
## - gender        1   516.16 542.16
##
## Step:  AIC=469.73
## smoking ~ age_group + gender + wealth_index + RESIDENCE + transport +
##     school + education
##
##                Df Deviance    AIC
## - wealth_index  1   444.21 468.21
## - transport     1   444.47 468.47
## - RESIDENCE     1   445.06 469.06
## <none>              443.73 469.73
## - school        1   445.77 469.77
## - age_group     4   452.06 470.06
## - education     3   451.68 471.68
## - gender        1   519.74 543.74
##
## Step:  AIC=468.21
## smoking ~ age_group + gender + RESIDENCE + transport + school +
##     education
##
##             Df Deviance    AIC
## - transport  1   444.89 466.89
## <none>           444.21 468.21
## - age_group  4   452.22 468.22
## - RESIDENCE  1   446.28 468.28
## - school     1   446.29 468.29
## - education  3   454.92 472.92
## - gender     1   520.09 542.09
##
## Step:  AIC=466.89
## smoking ~ age_group + gender + RESIDENCE + school + education
##
##             Df Deviance    AIC
## - age_group  4   452.66 466.66
## <none>           444.89 466.89
## - RESIDENCE  1   446.93 466.93
## - school     1   447.29 467.29
## - education  3   455.55 471.55
## - gender     1   521.24 541.24
##
## Step:  AIC=466.66
## smoking ~ gender + RESIDENCE + school + education
##
##             Df Deviance    AIC
## - RESIDENCE  1   454.01 466.01
## <none>           452.66 466.66
## - school     1   455.02 467.02
## - education  3   464.05 472.05
## - gender     1   533.91 545.91
##
## Step:  AIC=466.01
## smoking ~ gender + school + education
```

```
##
##             Df Deviance    AIC
## <none>          454.01 466.01
## - school     1   456.25 466.25
## - education  3   466.31 472.31
## - gender     1   535.20 545.20
```

```
##
## Call:  glm(formula = smoking ~ gender + school + education, family = "binomial",
##     data = new_df)
##
## Coefficients:
## (Intercept)      gender2       school  education.L  education.Q  education.C
##    -0.22707     -4.40263      0.55615     -1.12922      0.39884     -0.04688
##
## Degrees of Freedom: 450 Total (i.e. Null);  445 Residual
## Null Deviance:       545.3
## Residual Deviance: 454   AIC: 466
```

```r
####Avergae Marginal effect
###install.packages("margins")
library(margins)

# Calculating marginal effects
margins_obj <- margins(m1)

# Display summary of marginal effects
summary(margins_obj)
```

```
##                     factor     AME     SE        z      p   lower   upper
##             age_group35-49  0.0950 0.0481   1.9758 0.0482  0.0008  0.1892
##             age_group50-64  0.1071 0.0553   1.9365 0.0528 -0.0013  0.2154
##             age_group65-79 -0.0132 0.0872  -0.1515 0.8796 -0.1841  0.1577
##            age_group80-105 -0.2292 0.0356  -6.4377 0.0000 -0.2990 -0.1594
##    Division_NameChittagong  0.0266 0.0816   0.3257 0.7446 -0.1334  0.1865
##         Division_NameDhaka  0.1754 0.0834   2.1028 0.0355  0.0119  0.3389
##        Division_NameKhulna  0.1004 0.0688   1.4598 0.1443 -0.0344  0.2352
##    Division_NameMymensingh  0.1632 0.0738   2.2113 0.0270  0.0185  0.3078
##      Division_NameRajshahi  0.1273 0.0852   1.4944 0.1351 -0.0397  0.2943
##       Division_NameRangpur  0.1409 0.0709   1.9872 0.0469  0.0019  0.2798
##        Division_NameSylhet  0.1031 0.0777   1.3276 0.1843 -0.0491  0.2553
##    educationatbest_Primary -0.1820 0.0933  -1.9505 0.0511 -0.3648  0.0009
##  educationatbest_Secondary -0.2540 0.0879  -2.8912 0.0038 -0.4262 -0.0818
##  educationHigher_Education -0.2678 0.0888  -3.0169 0.0026 -0.4418 -0.0938
##                    gender2 -0.3729 0.0273 -13.6475 0.0000 -0.4264 -0.3193
##                   hospital  0.0277 0.0585   0.4746 0.6351 -0.0868  0.1423
##                 RESIDENCE2  0.0435 0.0455   0.9574 0.3384 -0.0456  0.1327
##                 restaurant -0.0060 0.0421  -0.1414 0.8875 -0.0885  0.0766
##                     school  0.1117 0.0662   1.6868 0.0916 -0.0181  0.2415
##                  transport  0.0310 0.0407   0.7616 0.4463 -0.0488  0.1108
##               wealth_index -0.0149 0.0275  -0.5415 0.5881 -0.0687  0.0390
```

```r
m_1 <- glm(smoking ~ age_group + gender + wealth_index + RESIDENCE + transport +
    school + education,data = new_df, family = "binomial")
summary(m_1)
```

```
##
## Call:
## glm(formula = smoking ~ age_group + gender + wealth_index + RESIDENCE +
##     transport + school + education, family = "binomial", data = new_df)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.77924    0.31795  -2.451  0.01425 *
## age_group35-49    0.55689    0.29535   1.886  0.05936 .
## age_group50-64    0.65719    0.33062   1.988  0.04684 *
## age_group65-79   -0.10642    0.59596  -0.179  0.85827
## age_group80-105 -14.04697  678.48404  -0.021  0.98348
## gender2          -4.36811    1.02691  -4.254  2.1e-05 ***
## wealth_index     -0.10354    0.14992  -0.691  0.48979
## RESIDENCE2        0.30066    0.26032   1.155  0.24811
## transport         0.19953    0.23271   0.857  0.39120
## school            0.54373    0.37960   1.432  0.15203
## education.L      -1.01769    0.37529  -2.712  0.00669 **
## education.Q       0.46274    0.30906   1.497  0.13432
## education.C      -0.07265    0.24986  -0.291  0.77123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 545.29  on 450  degrees of freedom
## Residual deviance: 443.73  on 438  degrees of freedom
## AIC: 469.73
##
## Number of Fisher Scoring iterations: 14
```

```r
library(lme4)

# 'strata' will be used as a random effect
model <- glmer(smoking ~ HH4B01 + gender + wealth_index + RESIDENCE + transport +
               school + hospital + restaurant + education + (1|gatsstrata),
               family = binomial, data = new_df)
```

```
## boundary (singular) fit: see help('isSingular')
```

```r
# Summary of the model
summary(model)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: smoking ~ HH4B01 + gender + wealth_index + RESIDENCE + transport +
##     school + hospital + restaurant + education + (1 | gatsstrata)
```

```
##    Data: new_df
##
##      AIC      BIC   logLik deviance df.resid
##    477.2    530.6   -225.6    451.2      438
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8828 -0.7064 -0.1336  0.8946  6.4729
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  gatsstrata (Intercept) 0        0
## Number of obs: 451, groups:  gatsstrata, 16
##
## Fixed effects:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.560106   0.457969  -1.223  0.22132
## HH4B01        0.005816   0.008627   0.674  0.50021
## gender2      -4.423431   1.027029  -4.307 1.65e-05 ***
## wealth_index -0.084398   0.148475  -0.568  0.56974
## RESIDENCE2    0.224809   0.254056   0.885  0.37622
## transport     0.171890   0.236416   0.727  0.46719
## school        0.531722   0.378151   1.406  0.15969
## hospital      0.143385   0.336726   0.426  0.67024
## restaurant   -0.143661   0.241127  -0.596  0.55132
## education.L  -1.015628   0.368677  -2.755  0.00587 **
## education.Q   0.396131   0.304085   1.303  0.19268
## education.C  -0.043606   0.246882  -0.177  0.85980
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) HH4B01 gendr2 wlth_n RESIDE trnspr school hosptl rstrnt
## HH4B01      -0.832
## gender2     -0.115  0.040
## wealth_indx -0.165 -0.072  0.029
## RESIDENCE2  -0.216 -0.038 -0.002  0.328
## transport   -0.182  0.020  0.000 -0.054 -0.014
## school      -0.088  0.037  0.025  0.036  0.061 -0.092
## hospital     0.005  0.020 -0.032 -0.139 -0.105 -0.128 -0.129
## restaurant  -0.208 -0.054  0.080  0.174  0.023 -0.210 -0.019 -0.110
## education.L -0.232  0.110  0.116 -0.331 -0.048  0.018 -0.029 -0.009 -0.034
## education.Q  0.196 -0.075 -0.084 -0.124 -0.017 -0.012  0.093 -0.004  0.065
## education.C -0.102  0.033  0.045  0.126  0.043  0.017  0.048  0.017 -0.060
##             edct.L edct.Q
## HH4B01
## gender2
## wealth_indx
## RESIDENCE2
## transport
## school
## hospital
## restaurant
## education.L
```

```
## education.Q -0.506
## education.C  0.200 -0.419
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

# Results

Our analysis utilized a GLM to investigate the determinants of smoking behavior among adults in Bangladesh. The model incorporated various socio-demographic and environmental factors, yielding insights into their respective influences on the likelihood of tobacco use.

The coefficient of age_group.L suggests a change in the log odds of smoking for this age group compared to the reference category, but the large standard errors and non-significant p-values indicate high uncertainty in this estimate. The coefficient age_group.Q reflects the quadratic relationship with smoking odds, yet the statistical insignificance suggests caution in interpretation. The rest of the age groups show non-significant associations with smoking behavior, as indicated by their large standard errors and p-values.

A coefficient of 4.46070 for gender indicates a substantial increase in the log odds of smoking for the second gender category compared to the first. This effect is statistically significant at the 0.001 level, highlighting gender as a significant predictor of smoking behavior.

Wealth index is not statistically significant, suggesting wealth status does not significantly influence smoking odds within this dataset. The variables residence, transport, hospital, restaurant did not show a statistically significant effect on the log odds of smoking. School showed marginally significant at the 0.1 level, indicating a potential influence of educational environments on smoking behavior.

The coefficients for education for different levels of education reveal that the linear term (education.L) is statistically significant at the 0.05 level, suggesting a notable difference in smoking odds between the reference education level and this particular level.

The coefficients for divisions like Dhaka and Mymensingh are statistically significant, indicating regional disparities in smoking behavior compared to the reference division.

The AIC value and Dharma residual plots indicate a good fit for the model, with residuals displaying an appropriate distribution.

While gender and education level show significant associations with smoking behavior, other variables do not exhibit statistically significant effects. The interpretation of these results should be approached with caution, particularly for coefficients with large standard errors and non-significant p-values.

So, our findings reveal that gender, education level, and geographical division are influential factors in smoking behavior among adults in Bangladesh. The model indicates that these variables have a statistically significant association with the likelihood of an individual being a smoker, suggesting that targeted interventions may be necessary to address these specific demographics.

The AIC value for our GLM was calculated to be 479.58, which provides a measure of the relative quality of the statistical model for a given set of data. This value aids in model selection, where a lower AIC value would indicate a better fit of the model to the data.

Further analysis was conducted using Dharma residual plots, which demonstrated that the residuals from our GLM had a good fit. This suggests that the model assumptions were met, and the variance was consistently distributed, indicating the reliability of our model's predictions.

Despite the presence of data clustering within our dataset, which often necessitates the use of Generalized Linear Mixed Models (GLMM) to account for random effects, our analysis revealed no significant random effects. This finding implies that the fixed effects model provided by GLM was sufficient for our data structure and the variables of interest.

Interestingly, the wealth index did not emerge as a statistically significant predictor in our model. This outcome suggests that, within the context of our study, wealth status did not have a discernible impact on smoking behavior among the surveyed population.

# Impact and Limitations

## Impact

*Gender, Education Level, and Location-based Impact:* The GLM model highlighting gender, education level, and division as influential factors underscores the need for any specific strategies. Implementing gender-sensitive smoking cessation programs and educational campaigns can have significant impact on society as addressing educational disparities can lead to better outcomes.

*Behavioral Interventions:* We can use these findings to design evidence-based interventions. For example, targeting low-income individuals, promoting smoke-free public spaces, and providing educational resources. Allocating resources to effective behavioral interventions and monitoring their impact can bring positive social outcomes.

## Limitations

*Lack of Causality:* Since our dataset captures only cross-sectional data at a specific point in time, they may reveal associations, but they cannot establish causality. In fact, our logistic regression provides associations but not causal relationships. We have identified associations between various factors and smoking, but causality will require rigorous experimental designs.

*Self-Reported Data:* GATS relies on self-reported information, which can be subject to recall bias or social desirability bias. Participants may underreport or overreport smoking behavior which would heavily influence our results.

*Missing Data Handling:* Dealing with missing values in time series can be tricky. Traditional methods like linear interpolation may not suffice, especially if the missingness is non-random.

# Conclusion

## Summary of Findings

*RQ1:* What are the socio-economic factors associated with tobacco smoking among adults in Bangladesh?

We found that gender, education level, and division (location) has an impact on smoking behavior among adults in Bangladesh.

*RQ2:* What is the impact of smoking tobacco in public places on the prevalence of smoking behavior of individuals in Bangladesh?

We were unable to prove that smoking tobacco in public places has any impact on smoking behavior of individuals in Bangladesh.

## Recommendations

*Causal Inference in Socio-Economic Studies:* There is scope to develop research designs beyond cross-sectional studies to establish causal relationships between socio-economic factors and smoking behavior. Longitudinal or experimental designs can be explored to establish causality here. However, policy-wise, funding and support for research is needed that focuses on causal inference, allowing for evidence-based policy decisions.

*Longitudinal Trajectories and Behavioral Memory:* Sophisticated machine learning approaches can be taken to model longitudinal trajectories using time series data. Addressing challenges such as missing data, synthetic data generation, and uncertainty estimation can help in this regard. The temporal patterns of smoking behavior over extended periods can also be investigated to understand how smoking habits evolve, including factors that trigger transitions (e.g., quitting attempts, relapses).

# References


1. Bandura, A. (2014). Social-cognitive theory. In An introduction to theories of personality (pp. 341-360). Psychology Press.

2. Hoque, M. M., & Tama, R. A. Z. (2021). Implementation of Tobacco Control Policies in Bangladesh: A Political Economy Analysis. Public Administration Research, 10(2), 36. https://doi.org/10.5539/par.v10n2p36

3. Hoque, Md Mahmudul & Tama, Riffat Ara Zannat (2021). Implementation of Tobacco Control Policies in Bangladesh: A Political Economy Analysis. Public Administration Research 10 (2):36-51. https://philpapers.org/rec/HOQIOT

4. Nargis, N., Faruque, G. M., Ahmed, M., Huq, I., Parven, R., Wadood, S. N., Hussain, A. G., & Drope, J. (2022). A comprehensive economic assessment of the health effects of tobacco use and implications for tobacco control in Bangladesh. Tobacco control, 31(6), 723–729. https://doi.org/10.1136/tobaccocontrol-2020-056175

5. Rosenstock, I. M. (1974). The Health Belief Model and Preventive Health Behavior. Health Education Monographs, 2(4), 354-386. https://journals.sagepub.com/doi/abs/10.1177/109019817400200405

6. Tobacco Control Research Group. (2023, September 14). Bangladesh Country Profile. Tobacco Tactics. Retrieved February 14, 2024, from https://tobaccotactics.org/article/bangladesh-country-profile/