Understanding Child Marriage Through In Vietnam’s Socio-Demographic Prism

Holly Duong

2024-02-29

# 1. Set Up

### Import Packages

# load packages  
library(readr)  
library(dplyr)  
library(ggplot2)  
library(gridExtra)  
library(corrplot)  
library(plotly)  
library(reshape2)  
library(car)  
library(kableExtra)  
library(broom)  
library(knitr)  
library(pROC)  
library(ggpattern)  
library(tidyr)  
library(ResourceSelection)  
library(sf)  
library(arm)  
library(xtable)  
  
# display packages versions  
#sessionInfo()

### Import Data

# import dataset  
female <- read\_csv("/Users/hollyduong/Desktop/DA 401/ChildMarriageInVietnam/Raw Data/wm.csv")

# get a glimpse of dataset  
head(female)

## # A tibble: 6 × 458  
## HH1 HH2 LN WM1 WM2 WM3 WMINT WM4 WM5 WM6D WM6M WM6Y WM8  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 2 3 1 2 3 92 91 92 19 11 2020 2  
## 2 1 4 2 1 4 2 92 91 92 18 11 2020 1  
## 3 1 9 4 1 9 4 92 91 92 18 11 2020 1  
## 4 1 10 4 1 10 4 92 91 92 18 11 2020 2  
## 5 1 11 4 1 11 4 92 91 92 19 11 2020 2  
## 6 1 11 5 1 11 5 92 91 92 18 11 2020 2  
## # ℹ 445 more variables: WM9 <dbl>, WM17 <dbl>, WM7H <dbl>, WM7M <dbl>,  
## # WM10H <dbl>, WM10M <dbl>, WM11 <dbl>, WM12 <dbl>, WM13 <dbl>, WM14 <dbl>,  
## # WM15 <dbl>, WM22 <dbl>, WM23 <dbl>, WM24 <dbl>, WMHINT <dbl>, WMFIN <dbl>,  
## # WB3M <dbl>, WB3Y <dbl>, WB4 <dbl>, WB5 <dbl>, WB6A <dbl>, WB6B <dbl>,  
## # WB7 <dbl>, WB9 <lgl>, WB10A <lgl>, WB10B <lgl>, WB11 <lgl>, WB12A <lgl>,  
## # WB12B <lgl>, WB14 <dbl>, WB15 <dbl>, WB16 <dbl>, WB17 <dbl>, WB18 <dbl>,  
## # WB19A <chr>, WB19B <chr>, WB19C <chr>, WB19D <chr>, WB19E <chr>, …

# 2. Data Pre-processing

### Variables Selection

# Select the specified columns to create a new dataframe  
female\_df <- dplyr::select(female, WAGEM, MSTATUS, HH6, HH7, welevel, insurance, ethnicity, windex5, CP2, HA1, MT4, MT9, MT11)  
  
# View the first few rows of the new dataframe  
summary(female\_df)

## WAGEM MSTATUS HH6 HH7 welevel   
## Min. :10.00 Min. :1.000 Min. :1.000 Min. :1.000 Min. :0.00   
## 1st Qu.:18.00 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:1.00   
## Median :20.00 Median :1.000 Median :2.000 Median :3.000 Median :2.00   
## Mean :20.92 Mean :1.413 Mean :1.683 Mean :3.404 Mean :2.46   
## 3rd Qu.:23.00 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:5.000 3rd Qu.:3.00   
## Max. :47.00 Max. :9.000 Max. :2.000 Max. :6.000 Max. :9.00   
## NA's :2026 NA's :11 NA's :11 NA's :11 NA's :524   
## insurance ethnicity windex5 CP2   
## Min. :1.000 Min. :1.000 Min. :0.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000   
## Median :1.000 Median :1.000 Median :2.000 Median :1.000   
## Mean :1.135 Mean :2.035 Mean :2.494 Mean :1.431   
## 3rd Qu.:1.000 3rd Qu.:3.000 3rd Qu.:4.000 3rd Qu.:2.000   
## Max. :9.000 Max. :6.000 Max. :5.000 Max. :9.000   
## NA's :524 NA's :864   
## HA1 MT4 MT9 MT11   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.00   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.00   
## Median :1.000 Median :2.000 Median :1.000 Median :1.00   
## Mean :1.215 Mean :1.664 Mean :1.365 Mean :1.12   
## 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:1.00   
## Max. :9.000 Max. :9.000 Max. :9.000 Max. :9.00   
## NA's :524 NA's :524 NA's :2496 NA's :524

# Rename the columns  
female\_df <- female\_df %>%  
 rename(  
 age\_first\_marriage = WAGEM,  
 marital\_status = MSTATUS,  
 area = HH6,  
 region = HH7,  
 education\_level = welevel,  
 health\_insurance = insurance,  
 ethnicity = ethnicity,  
 wealth\_index = windex5,  
 current\_contraceptive\_use = CP2,  
 awareness\_hiv\_aids = HA1,  
 used\_computer\_tablet = MT4,  
 used\_internet = MT9,  
 owns\_mobile\_phone = MT11  
)

### Check For Missing Data

# Summarize missing values by column  
summarize\_missing <- sapply(female\_df, function(x) sum(is.na(x)))  
print(summarize\_missing)

## age\_first\_marriage marital\_status area   
## 2026 11 11   
## region education\_level health\_insurance   
## 11 524 524   
## ethnicity wealth\_index current\_contraceptive\_use   
## 0 0 864   
## awareness\_hiv\_aids used\_computer\_tablet used\_internet   
## 524 524 2496   
## owns\_mobile\_phone   
## 524

### Data Cleaning

# Recode the value 9 to NA for specified variables  
female\_df <- female\_df %>%  
 mutate(  
 current\_contraceptive\_use = na\_if(current\_contraceptive\_use, 9),  
 used\_internet = na\_if(used\_internet, 9),  
 health\_insurance = na\_if(health\_insurance, 9),  
 education\_level = na\_if(education\_level, 9),  
 awareness\_hiv\_aids = na\_if(awareness\_hiv\_aids, 9),  
 used\_computer\_tablet = na\_if(used\_computer\_tablet, 9),  
 owns\_mobile\_phone = na\_if(owns\_mobile\_phone, 9),  
 marital\_status = na\_if(marital\_status, 9)  
 )  
  
# Recode the value 6 to NA for 'ethnicity'  
female\_df$ethnicity <- na\_if(female\_df$ethnicity, 6)  
  
# Recode the value 0 to NA for 'wealth\_index'  
female\_df$wealth\_index <- na\_if(female\_df$wealth\_index, 0)

# Recoding variables from 1 (Yes) and 2 (No) to 1 (Yes) and 0 (No)  
female\_df$health\_insurance <- ifelse(female\_df$health\_insurance == 2, 0, female\_df$health\_insurance)  
female\_df$current\_contraceptive\_use <- ifelse(female\_df$current\_contraceptive\_use == 2, 0, female\_df$current\_contraceptive\_use)  
female\_df$awareness\_hiv\_aids <- ifelse(female\_df$awareness\_hiv\_aids == 2, 0, female\_df$awareness\_hiv\_aids)  
female\_df$used\_computer\_tablet <- ifelse(female\_df$used\_computer\_tablet == 2, 0, female\_df$used\_computer\_tablet)  
female\_df$owns\_mobile\_phone <- ifelse(female\_df$owns\_mobile\_phone == 2, 0, female\_df$owns\_mobile\_phone)  
female\_df$used\_internet <- ifelse(female\_df$used\_internet == 2, 0, female\_df$used\_internet)

# View the changes to ensure NA substitution has been correctly applied  
summary(female\_df)

## age\_first\_marriage marital\_status area region   
## Min. :10.00 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:18.00 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000   
## Median :20.00 Median :1.000 Median :2.000 Median :3.000   
## Mean :20.92 Mean :1.408 Mean :1.683 Mean :3.404   
## 3rd Qu.:23.00 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:5.000   
## Max. :47.00 Max. :3.000 Max. :2.000 Max. :6.000   
## NA's :2026 NA's :18 NA's :11 NA's :11   
## education\_level health\_insurance ethnicity wealth\_index   
## Min. :0.00 Min. :0.0000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.00 1st Qu.:1.0000 1st Qu.:1.000 1st Qu.:1.000   
## Median :2.00 Median :1.0000 Median :1.000 Median :2.000   
## Mean :2.46 Mean :0.8659 Mean :1.586 Mean :2.615   
## 3rd Qu.:3.00 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:4.000   
## Max. :5.00 Max. :1.0000 Max. :4.000 Max. :5.000   
## NA's :525 NA's :525 NA's :1148 NA's :524   
## current\_contraceptive\_use awareness\_hiv\_aids used\_computer\_tablet  
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.0000   
## Median :1.0000 Median :1.0000 Median :0.0000   
## Mean :0.5842 Mean :0.7975 Mean :0.3412   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000   
## NA's :885 NA's :541 NA's :532   
## used\_internet owns\_mobile\_phone  
## Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.0000   
## Median :1.0000 Median :1.0000   
## Mean :0.6394 Mean :0.8831   
## 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000   
## NA's :2501 NA's :529

### Create New Variable Access to Media

# Combine individual variables for access to the internet, phone, and computer into a single variable  
# This new variable "access\_to\_media" will have a value of 1 if the respondent has access to any of these media sources, and 0 if not  
# This provides a more comprehensive measure of media access  
female\_df <- female\_df %>%  
 mutate(access\_to\_media = ifelse(used\_computer\_tablet == 1 | used\_internet == 1 | owns\_mobile\_phone == 1, 1, 0))

# In later analysis, use access\_to\_media instead of the 3 separate variables  
female\_df <- female\_df %>%  
 dplyr::select(-used\_computer\_tablet, -used\_internet, -owns\_mobile\_phone)

### Data Filtering (IMPORTANT: keeping only data of married demographics)

# Removing rows where 'age\_first\_marriage' is NA in female\_df  
# This is because these values represent people who are not yet married, which is not the focus of the study  
married\_df <- female\_df[!is.na(female\_df$age\_first\_marriage), ]  
summary(married\_df)

## age\_first\_marriage marital\_status area region   
## Min. :10.00 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:18.00 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000   
## Median :20.00 Median :1.000 Median :2.000 Median :3.000   
## Mean :20.92 Mean :1.063 Mean :1.708 Mean :3.377   
## 3rd Qu.:23.00 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:5.000   
## Max. :47.00 Max. :2.000 Max. :2.000 Max. :6.000   
##   
## education\_level health\_insurance ethnicity wealth\_index   
## Min. :0.00 Min. :0.0000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.00 1st Qu.:1.0000 1st Qu.:1.000 1st Qu.:1.000   
## Median :2.00 Median :1.0000 Median :1.000 Median :2.000   
## Mean :2.29 Mean :0.8597 Mean :1.641 Mean :2.541   
## 3rd Qu.:3.00 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:4.000   
## Max. :5.00 Max. :1.0000 Max. :4.000 Max. :5.000   
## NA's :405 NA's :404 NA's :980 NA's :404   
## current\_contraceptive\_use awareness\_hiv\_aids access\_to\_media   
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:1.0000   
## Median :1.0000 Median :1.0000 Median :1.0000   
## Mean :0.7099 Mean :0.7756 Mean :0.8918   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000   
## NA's :744 NA's :415 NA's :408

# Exporting female\_df to a CSV file in the current working directory  
#write.csv(married\_df, "married\_df.csv", row.names = FALSE)

### Missing Data Imputation (For Approach 1)

# Before that, let's double check the count of missing values by column  
count\_missing <- sapply(married\_df, function(x) sum(is.na(x)))  
print(count\_missing)

## age\_first\_marriage marital\_status area   
## 0 0 0   
## region education\_level health\_insurance   
## 0 405 404   
## ethnicity wealth\_index current\_contraceptive\_use   
## 980 404 744   
## awareness\_hiv\_aids access\_to\_media   
## 415 408

# Make a copy of female\_df for imputation  
imputed\_df <- married\_df

# Define a function to calculate mode for categorical variables  
getMode <- function(v) {  
 # The mode is the value that appears most frequently in the data  
 uniqv <- unique(na.omit(v)) # Omit NA values and get unique values  
 uniqv[which.max(tabulate(match(v, uniqv)))] # Return the value with the highest frequency  
}

# For 'ethnicity', an ordinal variable with predefined categories, it makes sense to impute missing values with the mode.  
mode\_ethnicity <- getMode(imputed\_df$ethnicity)  
imputed\_df <- mutate(imputed\_df, ethnicity = ifelse(is.na(ethnicity), mode\_ethnicity, ethnicity))  
mode\_area <- getMode(imputed\_df$area)  
imputed\_df <- mutate(imputed\_df, area = ifelse(is.na(area), mode\_area, area))  
mode\_region <- getMode(imputed\_df$region)  
imputed\_df <- mutate(imputed\_df, region = ifelse(is.na(region), mode\_region, region))  
mode\_marital\_status <- getMode(imputed\_df$marital\_status)  
imputed\_df <- mutate(imputed\_df, marital\_status = ifelse(is.na(marital\_status), mode\_marital\_status, marital\_status))  
  
  
# Binary variables like 'current\_contraceptive\_use', 'health\_insurance','awareness\_hiv\_aids', 'used\_internet', 'used\_computer\_tablet', 'owns\_mobile\_phone', and 'access\_to\_media'  
# should be imputed with the mode since it represents the most frequent category (either 0 or 1).  
  
# Calculate the mode for each binary variable  
mode\_current\_contraceptive\_use <- getMode(imputed\_df$current\_contraceptive\_use)  
mode\_health\_insurance <- getMode(imputed\_df$health\_insurance)  
mode\_awareness\_hiv\_aids <- getMode(imputed\_df$awareness\_hiv\_aids)  
mode\_access\_to\_media <- getMode(imputed\_df$access\_to\_media)  
  
# Impute missing values for binary variables  
imputed\_df <- mutate(imputed\_df,  
 current\_contraceptive\_use = ifelse(is.na(current\_contraceptive\_use), mode\_current\_contraceptive\_use, current\_contraceptive\_use),  
 health\_insurance = ifelse(is.na(health\_insurance), mode\_health\_insurance, health\_insurance),  
 awareness\_hiv\_aids = ifelse(is.na(awareness\_hiv\_aids), mode\_awareness\_hiv\_aids, awareness\_hiv\_aids),  
 access\_to\_media = ifelse(is.na(access\_to\_media), mode\_access\_to\_media, access\_to\_media)  
)  
  
# 'education\_level' is an ordinal variable where the median could be a more suitable measure of central tendency than the mode.  
# However, given the categorical nature of the levels (e.g., "Primary", "Secondary"), using the mode may still be appropriate.  
mode\_education\_level <- getMode(imputed\_df$education\_level)  
imputed\_df <- mutate(imputed\_df, education\_level = ifelse(is.na(education\_level), mode\_education\_level, education\_level))  
  
# 'wealth\_index' is an ordinal variable where the median could be a more suitable measure of central tendency than the mode.  
# However, given the categorical nature of the levels (e.g., "Poorest", "Poor",...), using the mode may still be appropriate.  
mode\_wealth\_index <- getMode(imputed\_df$wealth\_index)  
imputed\_df <- mutate(imputed\_df, wealth\_index = ifelse(is.na(wealth\_index), mode\_wealth\_index, wealth\_index))

# Check the resulting dataset to confirm changes  
summary(imputed\_df)

## age\_first\_marriage marital\_status area region   
## Min. :10.00 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:18.00 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000   
## Median :20.00 Median :1.000 Median :2.000 Median :3.000   
## Mean :20.92 Mean :1.063 Mean :1.708 Mean :3.377   
## 3rd Qu.:23.00 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:5.000   
## Max. :47.00 Max. :2.000 Max. :2.000 Max. :6.000   
## education\_level health\_insurance ethnicity wealth\_index   
## Min. :0.000 Min. :0.0000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:1.0000 1st Qu.:1.000 1st Qu.:1.000   
## Median :2.000 Median :1.0000 Median :1.000 Median :2.000   
## Mean :2.278 Mean :0.8658 Mean :1.574 Mean :2.474   
## 3rd Qu.:3.000 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:4.000   
## Max. :5.000 Max. :1.0000 Max. :4.000 Max. :5.000   
## current\_contraceptive\_use awareness\_hiv\_aids access\_to\_media   
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:1.0000   
## Median :1.0000 Median :1.0000 Median :1.0000   
## Mean :0.7332 Mean :0.7856 Mean :0.8965   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000

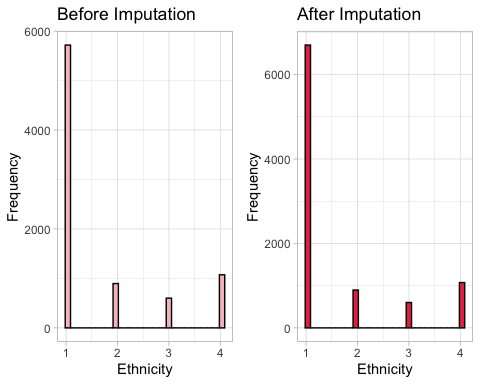
# After imputation, let's check for missing values  
count\_imputation <- sapply(imputed\_df, function(x) sum(is.na(x)))  
print(count\_imputation)

## age\_first\_marriage marital\_status area   
## 0 0 0   
## region education\_level health\_insurance   
## 0 0 0   
## ethnicity wealth\_index current\_contraceptive\_use   
## 0 0 0   
## awareness\_hiv\_aids access\_to\_media   
## 0 0

### Visualization Comparing Before vs. After Imputation

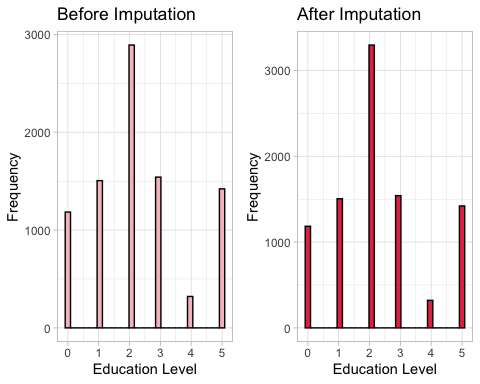
# Histogram for 'ethnicity' before imputation  
ethnic <- ggplot(married\_df, aes(x = ethnicity)) +   
 geom\_histogram(fill = "#F7C0C8", color = "black", bins = 30) +  
 theme\_light() +  
 ggtitle("Before Imputation") +  
 xlab("Ethnicity") +  
 ylab("Frequency")  
  
# Histogram for 'ethnicity' after imputation  
ethnic\_imputed <- ggplot(imputed\_df, aes(x = ethnicity)) +   
 geom\_histogram(fill = "#E83853", color = "black", bins = 30) +  
 theme\_light() +  
 ggtitle("After Imputation") +  
 xlab("Ethnicity") +  
 ylab("Frequency")  
  
# Arrange the two plots side by side  
grid.arrange(ethnic, ethnic\_imputed, ncol = 2)

## Warning: Removed 980 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



# Histogram for 'education\_level' before imputation  
educ <- ggplot(married\_df, aes(x = education\_level)) +   
 geom\_histogram(fill = "#F7C0C8", color = "black", bins = 30) +  
 theme\_light() +  
 ggtitle("Before Imputation") +  
 xlab("Education Level") +  
 ylab("Frequency")  
  
# Histogram for 'education\_level' after imputation  
educ\_imputed <- ggplot(imputed\_df, aes(x = education\_level)) +   
 geom\_histogram(fill = "#E83853", color = "black", bins = 30) +  
 theme\_light() +  
 ggtitle("After Imputation") +  
 xlab("Education Level") +  
 ylab("Frequency")  
  
# Arrange the two plots side by side  
grid.arrange(educ, educ\_imputed, ncol = 2)

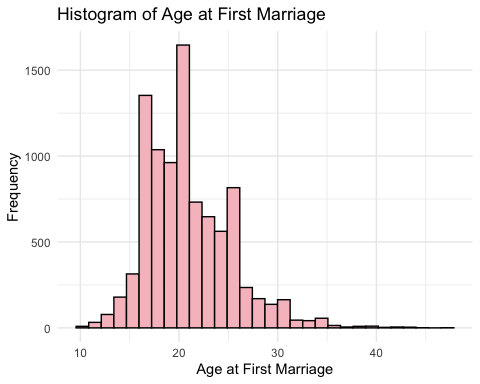
## Warning: Removed 405 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



### Age\_first\_marriage Distribution

# Histogram for 'Age at First Marriage'  
afm\_hist <- ggplot(imputed\_df, aes(x = age\_first\_marriage)) +  
 geom\_histogram(fill = "#F7C0C8", color = "black") +  
 theme\_minimal() +  
 ggtitle("Histogram of Age at First Marriage") +  
 xlab("Age at First Marriage") +  
 ylab("Frequency")  
  
print(afm\_hist)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#ggsave("age\_first\_marriage.png", plot = afm\_hist, width = 8, height = 6)

### Creating Binary Variables for Child Marriage Under 18 and 16

# Convert "age at first marriage" into a binary variable to indicate child marriage  
# Child marriage is defined as marriage before the age of 18  
# The new binary variable "child\_marriage" will have a value of 1 if the marriage occurred before age 18, and 0 otherwise  
imputed\_df <- imputed\_df %>%  
 mutate(child\_marriage = ifelse(age\_first\_marriage < 18, 1, 0))

# Create a binary variable for child marriage under 16  
# The new variable "child\_marriage\_u16" will have a value of 1 if the marriage occurred before age 16, and 0 otherwise  
imputed\_df <- imputed\_df %>%  
 mutate(child\_marriage\_u16 = ifelse(age\_first\_marriage < 16, 1, 0))

# Move "child\_marriage" and "child\_marriage\_u16" to the front of the dataframe  
imputed\_df <- imputed\_df %>%  
 dplyr::select(child\_marriage, child\_marriage\_u16, everything())

# Exporting female\_df to a CSV file in the current working directory  
#write.csv(imputed\_df, "imputed\_df.csv", row.names = FALSE)

# 3. Exploratory Data Analysis

### Vietnam Map with Regions and Their Percentages of Child Marriage (under 18)

# Aggregate the data by region to get the total number of child marriages under 18 per region  
region\_counts <- aggregate(child\_marriage ~ region, data = imputed\_df, FUN = sum)  
  
# Calculate the total number of child marriages under 18 in the dataset  
total\_child\_marriages <- sum(region\_counts$child\_marriage)  
  
# Calculate the percentage for each region  
region\_counts$married\_u18\_perc\_of\_total <- (region\_counts$child\_marriage / total\_child\_marriages) \* 100

# Mapping region numbers to names  
region\_names <- c("Red River Delta", "Northern Midlands And Mountain",   
 "North Central And Central Coastal", "Central Highlands",   
 "South East", "Mekong River Delta")  
names(region\_counts)[1] <- "region\_name"  
region\_counts$region\_name <- factor(region\_counts$region\_name, levels = 1:6, labels = region\_names)  
  
# Display the final data frame  
print(region\_counts)

## region\_name child\_marriage married\_u18\_perc\_of\_total  
## 1 Red River Delta 142 7.226463  
## 2 Northern Midlands And Mountain 888 45.190840  
## 3 North Central And Central Coastal 215 10.941476  
## 4 Central Highlands 271 13.791349  
## 5 South East 194 9.872774  
## 6 Mekong River Delta 255 12.977099

# Updated mapping including all provinces and cities in the Red River Delta  
province\_to\_region <- data.frame(  
 NAME\_1 = c(  
 'Bắc Ninh', 'Hà Nam', 'Hà Nội', 'Hải Dương', 'Hải Phòng', 'Hoà Bình', 'Hưng Yên', 'Nam Định', 'Ninh Bình', 'Thái Bình', 'Vĩnh Phúc', # Red River Delta 11  
   
 'Bắc Giang', 'Bắc Kạn', 'Cao Bằng', 'Hà Giang', 'Lạng Sơn', 'Lào Cai', 'Phú Thọ', 'Quảng Ninh', 'Thái Nguyên', 'Tuyên Quang', 'Yên Bái', 'Điện Biên', 'Lai Châu', 'Sơn La', # Northern Midlands And Mountain 14  
   
 'Bình Định', 'Bình Thuận', 'Khánh Hòa', 'Ninh Thuận', 'Phú Yên', 'Quảng Nam', 'Quảng Ngãi', 'Thừa Thiên Huế', 'Đà Nẵng', 'Hà Tĩnh', 'Nghệ An', 'Quảng Bình', 'Quảng Trị', 'Thanh Hóa', # North Central And Central Coastal 14  
   
 'Đắk Lắk', 'Đắk Nông', 'Gia Lai', 'Kon Tum', 'Lâm Đồng', # Central Highlands 5  
   
 'Bà Rịa - Vũng Tàu', 'Bình Dương', 'Bình Phước', 'Đồng Nai', 'Hồ Chí Minh', 'Tây Ninh', # South East 6  
   
 'An Giang', 'Bạc Liêu', 'Bến Tre', 'Cà Mau', 'Cần Thơ', 'Đồng Tháp', 'Hậu Giang', 'Kiên Giang', 'Long An', 'Sóc Trăng', 'Tiền Giang', 'Trà Vinh', 'Vĩnh Long' # Mekong River Delta 13  
 ),  
 Region = c(  
 rep('Red River Delta', 11),   
 rep('Northern Midlands And Mountain', 14),   
 rep('North Central And Central Coastal', 14),   
 rep('Central Highlands', 5),   
 rep('South East', 6),   
 rep('Mekong River Delta', 13)  
 )  
)

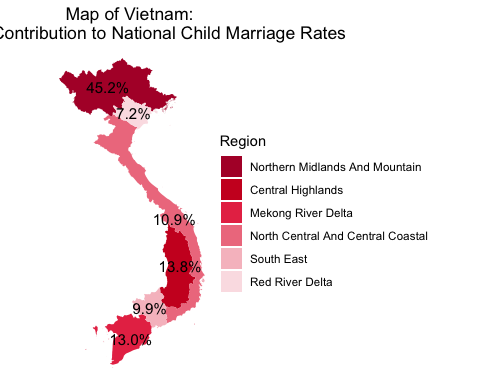
# Read shapefile  
vietnam\_shape <- st\_read('/Users/hollyduong/Desktop/DA 401/ChildMarriageInVietnam/gadm41\_VNM\_shp')

## Multiple layers are present in data source /Users/hollyduong/Desktop/DA 401/ChildMarriageInVietnam/gadm41\_VNM\_shp, reading layer `gadm41\_VNM\_1'.  
## Use `st\_layers' to list all layer names and their type in a data source.  
## Set the `layer' argument in `st\_read' to read a particular layer.  
## Reading layer `gadm41\_VNM\_1' from data source   
## `/Users/hollyduong/Desktop/DA 401/ChildMarriageInVietnam/gadm41\_VNM\_shp'   
## using driver `ESRI Shapefile'  
## Simple feature collection with 63 features and 11 fields  
## Geometry type: MULTIPOLYGON  
## Dimension: XY  
## Bounding box: xmin: 102.1446 ymin: 8.381355 xmax: 109.4692 ymax: 23.39269  
## Geodetic CRS: WGS 84

# Join the shapefile with the province-to-region mapping  
vietnam\_shape\_with\_region <- vietnam\_shape %>%  
 left\_join(province\_to\_region, by = "NAME\_1")  
  
# Aggregate the shapefile data by region  
vietnam\_regions <- vietnam\_shape\_with\_region %>%  
 group\_by(Region) %>%  
 summarise(geometry = st\_union(geometry), .groups = 'drop')

# Join the aggregated shapefile data with the child marriage data  
vietnam\_map\_data <- vietnam\_regions %>%  
 left\_join(region\_counts, by = c("Region" = "region\_name"))  
  
# Define colors with your specific choices  
colors\_ordered <- setNames(c("#B20033", "#CD0A25", "#E83853", "#EF7D8D", "#F7C0C8", "#FBE1E5"),  
 c("Northern Midlands And Mountain", "Central Highlands",   
 "Mekong River Delta", "North Central And Central Coastal",   
 "South East", "Red River Delta"))  
  
# Plot  
mapvn <- ggplot(data = vietnam\_map\_data) +  
 geom\_sf(aes(fill = factor(Region, levels = names(colors\_ordered))), color = NA) +  
 geom\_sf\_text(aes(label = sprintf("%.1f%%", married\_u18\_perc\_of\_total)), size = 4, hjust = 0.5, vjust = 0.5) +  
 scale\_fill\_manual(values = colors\_ordered, name = "Region") +  
 labs(title = "Map of Vietnam: \nRegional Contribution to National Child Marriage Rates") +  
 theme\_void() +  
 theme(legend.position = "right") +  
 theme(plot.title = element\_text(hjust = 0.5))  
  
print(mapvn)

## Warning in st\_point\_on\_surface.sfc(sf::st\_zm(x)): st\_point\_on\_surface may not  
## give correct results for longitude/latitude data



#ggsave("mapvn.png", plot = mapvn, width = 8, height = 6, dpi = 300)

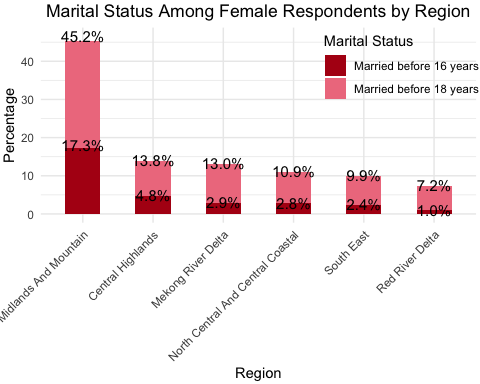
### Marital Status Among Female Respondants by Region in Vietnam

total\_child\_marriages\_u18 <- sum(imputed\_df$child\_marriage == 1)  
  
# Aggregate counts for each category by region  
counts\_df <- aggregate(cbind(married\_u18 = imputed\_df$child\_marriage == 1,   
 married\_u16 = imputed\_df$child\_marriage\_u16 == 1) ~ region,   
 data = imputed\_df,   
 FUN = sum)  
  
# Convert counts to percentages based on total child marriages under 18  
counts\_df$married\_u18 <- (counts\_df$married\_u18 / total\_child\_marriages\_u18) \* 100  
counts\_df$married\_u16 <- (counts\_df$married\_u16 / total\_child\_marriages\_u18) \* 100

# Add a new column for total percentages  
counts\_df$total\_percentage <- counts\_df$married\_u18 + counts\_df$married\_u16  
  
p <- ggplot(counts\_df, aes(x = reorder(factor(region), -total\_percentage))) +  
 geom\_bar(aes(y = married\_u18, fill = "Married before 18 years old"), stat = "identity", width = 0.5) +  
 geom\_bar(aes(y = married\_u16, fill = "Married before 16 years old"), stat = "identity", width = 0.5) +  
   
 # Adding text labels for married\_u18  
 geom\_text(aes(y = married\_u18, label = sprintf("%.1f%%", married\_u18)),   
 position = position\_stack(vjust = 1.03),   
 size = 4, color = "black") +  
 # Adding text labels for married\_u16  
 geom\_text(aes(y = married\_u16, label = sprintf("%.1f%%", married\_u16)),   
 position = position\_stack(vjust = 1.05),   
 size = 4, color = "black") +  
   
 scale\_fill\_manual(values = c("Married before 18 years old" = "#EF7D8D",   
 "Married before 16 years old" = "#B20016"),  
 name = "Marital Status") +  
 labs(x = "Region", y = "Percentage", title = "Marital Status Among Female Respondents by Region", element\_text(size = 14)) +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5),  
 plot.margin = margin(t = 1, r = 1, b = 1, l = 1, unit = "mm"),   
 axis.text.x = element\_text(angle = 45, hjust = 1),   
 legend.position = c(0.85,0.8))+  
 scale\_x\_discrete(labels = c("1" = "Red River Delta", "2" = "Northern Midlands And Mountain",   
 "3" = "North Central And Central Coastal", "4" = "Central Highlands",   
 "5" = "South East", "6" = "Mekong River Delta"))

## Warning: A numeric `legend.position` argument in `theme()` was deprecated in ggplot2  
## 3.5.0.  
## ℹ Please use the `legend.position.inside` argument of `theme()` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

print(p)



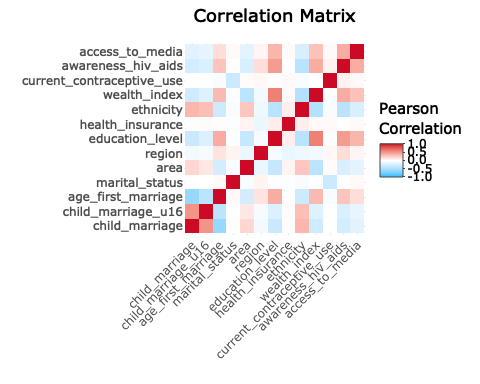
#ggsave("marital\_status.png", plot = p, width = 8, height = 7)

# 4. Pre-Analysis

### Correlation Matrix

# Calculate correlation matrix  
cor\_matrix <- cor(imputed\_df %>% select\_if(is.numeric), use = "complete.obs")  
  
# Melt the correlation matrix  
melted\_cor\_matrix <- melt(cor\_matrix)

# Generate an interactive heatmap  
corr\_matrix <- ggplot(melted\_cor\_matrix, aes(Var1, Var2, fill = value)) +  
 geom\_tile() +  
 scale\_fill\_gradientn(  
 colours = c("deepskyblue", "white", "#CD0A25"),  
 values = scales::rescale(c(-1, 0, 1)),  
 limits = c(-1, 1),  
 name="Pearson\nCorrelation"  
 ) +  
 theme\_minimal() +   
 theme(plot.title = element\_text(hjust = 0.5)) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 xlab("") +   
 ylab("") +  
 ggtitle("Correlation Matrix")   
  
# Convert ggplot object to plotly for interactivity  
ggplotly(corr\_matrix)



#write.csv(imputed\_df, "imputed\_df.csv", row.names = FALSE)

### Converting Categorical and Binary Variables to Factors

# Convert nominal and ordinal variables to factors  
imputed\_df$area <- as.factor(imputed\_df$area)  
imputed\_df$region <- as.factor(imputed\_df$region)  
imputed\_df$education\_level <- factor(imputed\_df$education\_level, ordered = FALSE)  
imputed\_df$ethnicity <- as.factor(imputed\_df$ethnicity)  
imputed\_df$wealth\_index <- factor(imputed\_df$wealth\_index, ordered = FALSE)  
  
# Binary variables are already in the correct format and can be used as is

# 5. Approach 1: Logistic Regression Analysis with Data Imputation

## Model 1 - Base Logistic Regression Model

# Baseline model for reference  
baseline\_model <- glm(child\_marriage ~ area + education\_level + wealth\_index + health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media, family = binomial(), data = imputed\_df)  
  
# Summarize the baseline model  
summary(baseline\_model)

##   
## Call:  
## glm(formula = child\_marriage ~ area + education\_level + wealth\_index +   
## health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids +   
## access\_to\_media, family = binomial(), data = imputed\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.35056 0.13405 -2.615 0.00892 \*\*   
## area2 0.40764 0.08124 5.018 5.22e-07 \*\*\*  
## education\_level1 -0.36537 0.08856 -4.126 3.70e-05 \*\*\*  
## education\_level2 -0.43403 0.08711 -4.983 6.27e-07 \*\*\*  
## education\_level3 -1.11556 0.11658 -9.569 < 2e-16 \*\*\*  
## education\_level4 -3.14435 0.51246 -6.136 8.47e-10 \*\*\*  
## education\_level5 -2.92930 0.25430 -11.519 < 2e-16 \*\*\*  
## wealth\_index2 -0.54988 0.08114 -6.777 1.23e-11 \*\*\*  
## wealth\_index3 -0.93401 0.10076 -9.270 < 2e-16 \*\*\*  
## wealth\_index4 -0.76067 0.11045 -6.887 5.70e-12 \*\*\*  
## wealth\_index5 -1.02057 0.15040 -6.786 1.16e-11 \*\*\*  
## health\_insurance 0.11657 0.08174 1.426 0.15384   
## current\_contraceptive\_use -0.03278 0.06298 -0.520 0.60280   
## awareness\_hiv\_aids -0.45273 0.07058 -6.415 1.41e-10 \*\*\*  
## access\_to\_media -0.02742 0.08280 -0.331 0.74051   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 9576.1 on 9267 degrees of freedom  
## Residual deviance: 8000.9 on 9253 degrees of freedom  
## AIC: 8030.9  
##   
## Number of Fisher Scoring iterations: 7

## Model 2 - Adding Fixed Effects to Base Model

# Enhanced model with additional fixed effects  
enhanced\_model <- glm(child\_marriage ~ area + region + education\_level + ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media,   
 family = binomial(),   
 data = imputed\_df)  
  
# Summarize the new model with FEs  
summary(enhanced\_model)

##   
## Call:  
## glm(formula = child\_marriage ~ area + region + education\_level +   
## ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media, family = binomial(),   
## data = imputed\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.193525 0.182449 -6.542 6.08e-11 \*\*\*  
## area2 0.268252 0.085814 3.126 0.00177 \*\*   
## region2 0.364102 0.130710 2.786 0.00534 \*\*   
## region3 0.162343 0.129158 1.257 0.20878   
## region4 0.331093 0.127873 2.589 0.00962 \*\*   
## region5 -0.009249 0.127505 -0.073 0.94217   
## region6 -0.041750 0.134893 -0.310 0.75694   
## education\_level1 0.023045 0.096277 0.239 0.81083   
## education\_level2 -0.031527 0.095643 -0.330 0.74167   
## education\_level3 -0.739558 0.124044 -5.962 2.49e-09 \*\*\*  
## education\_level4 -2.815602 0.514641 -5.471 4.47e-08 \*\*\*  
## education\_level5 -2.597688 0.257538 -10.087 < 2e-16 \*\*\*  
## ethnicity2 -0.004214 0.106644 -0.040 0.96848   
## ethnicity3 0.284476 0.130154 2.186 0.02884 \*   
## ethnicity4 1.200228 0.109158 10.995 < 2e-16 \*\*\*  
## wealth\_index2 -0.230751 0.086382 -2.671 0.00756 \*\*   
## wealth\_index3 -0.612377 0.106964 -5.725 1.03e-08 \*\*\*  
## wealth\_index4 -0.432981 0.118543 -3.653 0.00026 \*\*\*  
## wealth\_index5 -0.685336 0.162011 -4.230 2.33e-05 \*\*\*  
## health\_insurance 0.002808 0.083641 0.034 0.97321   
## current\_contraceptive\_use 0.031968 0.065069 0.491 0.62322   
## awareness\_hiv\_aids -0.184659 0.077076 -2.396 0.01658 \*   
## access\_to\_media -0.081214 0.087354 -0.930 0.35252   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 9576.1 on 9267 degrees of freedom  
## Residual deviance: 7749.7 on 9245 degrees of freedom  
## AIC: 7795.7  
##   
## Number of Fisher Scoring iterations: 7

## Model 3 - Model with Interaction Terms (area:education\_level)

# Adding the interaction term between area and education level  
model\_interaction <- update(enhanced\_model, . ~ . + area:education\_level)  
summary(model\_interaction)

##   
## Call:  
## glm(formula = child\_marriage ~ area + region + education\_level +   
## ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media + area:education\_level,   
## family = binomial(), data = imputed\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.805452 0.304616 -2.644 0.008189 \*\*   
## area2 -0.148776 0.275355 -0.540 0.588986   
## region2 0.353114 0.131039 2.695 0.007045 \*\*   
## region3 0.155871 0.129375 1.205 0.228281   
## region4 0.331470 0.128087 2.588 0.009658 \*\*   
## region5 0.003210 0.127831 0.025 0.979965   
## region6 -0.042289 0.134983 -0.313 0.754058   
## education\_level1 -0.154242 0.303997 -0.507 0.611889   
## education\_level2 -0.518954 0.283108 -1.833 0.066793 .   
## education\_level3 -1.183848 0.318697 -3.715 0.000203 \*\*\*  
## education\_level4 -3.614007 1.039464 -3.477 0.000507 \*\*\*  
## education\_level5 -2.978633 0.436539 -6.823 8.90e-12 \*\*\*  
## ethnicity2 -0.003071 0.106724 -0.029 0.977042   
## ethnicity3 0.268275 0.130228 2.060 0.039394 \*   
## ethnicity4 1.222054 0.110105 11.099 < 2e-16 \*\*\*  
## wealth\_index2 -0.243916 0.086709 -2.813 0.004907 \*\*   
## wealth\_index3 -0.621079 0.107275 -5.790 7.06e-09 \*\*\*  
## wealth\_index4 -0.432211 0.118897 -3.635 0.000278 \*\*\*  
## wealth\_index5 -0.679620 0.163475 -4.157 3.22e-05 \*\*\*  
## health\_insurance 0.009255 0.083784 0.110 0.912046   
## current\_contraceptive\_use 0.035227 0.065108 0.541 0.588473   
## awareness\_hiv\_aids -0.192308 0.077046 -2.496 0.012560 \*   
## access\_to\_media -0.085099 0.087235 -0.976 0.329304   
## area2:education\_level1 0.176156 0.318702 0.553 0.580448   
## area2:education\_level2 0.547315 0.295447 1.852 0.063954 .   
## area2:education\_level3 0.499730 0.338962 1.474 0.140402   
## area2:education\_level4 1.023762 1.194225 0.857 0.391301   
## area2:education\_level5 0.409921 0.541866 0.756 0.449350   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 9576.1 on 9267 degrees of freedom  
## Residual deviance: 7743.7 on 9240 degrees of freedom  
## AIC: 7799.7  
##   
## Number of Fisher Scoring iterations: 7

## Models Comparison (Odd Ratios and 95% Confidence Intervals)

# Function to add significance asterisks  
add\_asterisks <- function(p\_value) {  
 if (is.na(p\_value)) {  
 return(NA)  
 } else if (p\_value < 0.001) {  
 return("\*\*\*")  
 } else if (p\_value < 0.01) {  
 return("\*\*")  
 } else if (p\_value < 0.05) {  
 return("\*")  
 } else {  
 return("")  
 }  
}

# Function to format confidence intervals as a string  
format\_ci <- function(lower, upper) {  
 paste0("(", round(lower, 2), ", ", round(upper, 2), ")")  
}

# Tidy the baseline model with confidence intervals  
tidy\_baseline <- tidy(baseline\_model, conf.int = TRUE, exponentiate = TRUE)  
  
# Tidy the enhanced model with confidence intervals  
tidy\_enhanced <- tidy(enhanced\_model, conf.int = TRUE, exponentiate = TRUE)  
  
# Tidy the interaction model with confidence intervals  
tidy\_interaction <- tidy(model\_interaction, conf.int = TRUE, exponentiate = TRUE)

# Apply the function to each model's p.value  
tidy\_baseline$asterisks <- sapply(tidy\_baseline$p.value, add\_asterisks)  
tidy\_enhanced$asterisks <- sapply(tidy\_enhanced$p.value, add\_asterisks)  
tidy\_interaction$asterisks <- sapply(tidy\_interaction$p.value, add\_asterisks)

# Create OR strings with asterisks and format CIs as a string  
tidy\_baseline <- tidy\_baseline %>%  
 mutate(  
 OR = ifelse(is.na(estimate), NA, paste0(round(estimate, 2), asterisks)),  
 CI = ifelse(is.na(conf.low) | is.na(conf.high), NA, format\_ci(conf.low, conf.high))  
 )  
  
tidy\_enhanced <- tidy\_enhanced %>%  
 mutate(  
 OR = ifelse(is.na(estimate), NA, paste0(round(estimate, 2), asterisks)),  
 CI = ifelse(is.na(conf.low) | is.na(conf.high), NA, format\_ci(conf.low, conf.high))  
 )  
  
tidy\_interaction <- tidy\_interaction %>%  
 mutate(  
 OR = ifelse(is.na(estimate), NA, paste0(round(estimate, 2), asterisks)),  
 CI = ifelse(is.na(conf.low) | is.na(conf.high), NA, format\_ci(conf.low, conf.high))  
 )

# Add a 'Model' column to each tidied dataframe  
tidy\_baseline <- tidy\_baseline %>% mutate(Model = "Model\_1")  
tidy\_enhanced <- tidy\_enhanced %>% mutate(Model = "Model\_2")  
tidy\_interaction <- tidy\_interaction %>% mutate(Model = "Model\_3")  
  
# Combine and pivot the dataframes  
combined\_results <- bind\_rows(  
 tidy\_baseline %>% dplyr::select(term, OR, CI, Model),  
 tidy\_enhanced %>% dplyr::select(term, OR, CI, Model),  
 tidy\_interaction %>% dplyr::select(term, OR, CI, Model)  
) %>%  
 pivot\_wider(names\_from = Model, values\_from = c(OR, CI))

# Replace NAs with "—"  
combined\_results[is.na(combined\_results)] <- "—"

# Reordering columns to have OR and CI next to each other for each model  
combined\_results <- combined\_results %>%  
 dplyr::select(term,   
 OR\_Model\_1, CI\_Model\_1,   
 OR\_Model\_2, CI\_Model\_2,   
 OR\_Model\_3, CI\_Model\_3)

# Print the final combined table  
print(combined\_results)

## # A tibble: 28 × 7  
## term OR\_Model\_1 CI\_Model\_1 OR\_Model\_2 CI\_Model\_2 OR\_Model\_3 CI\_Model\_3  
## <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 (Intercept) 0.7\*\* (0.54, 0.… 0.3\*\*\* (0.21, 0.… 0.45\*\* (0.24, 0.…  
## 2 area2 1.5\*\*\* (1.28, 1.… 1.31\*\* (1.11, 1.… 0.86 (0.51, 1.…  
## 3 education\_… 0.69\*\*\* (0.58, 0.… 1.02 (0.85, 1.… 0.86 (0.48, 1.…  
## 4 education\_… 0.65\*\*\* (0.55, 0.… 0.97 (0.8, 1.1… 0.6 (0.35, 1.…  
## 5 education\_… 0.33\*\*\* (0.26, 0.… 0.48\*\*\* (0.37, 0.… 0.31\*\*\* (0.16, 0.…  
## 6 education\_… 0.04\*\*\* (0.01, 0.… 0.06\*\*\* (0.02, 0.… 0.03\*\*\* (0, 0.14)   
## 7 education\_… 0.05\*\*\* (0.03, 0.… 0.07\*\*\* (0.04, 0.… 0.05\*\*\* (0.02, 0.…  
## 8 wealth\_ind… 0.58\*\*\* (0.49, 0.… 0.79\*\* (0.67, 0.… 0.78\*\* (0.66, 0.…  
## 9 wealth\_ind… 0.39\*\*\* (0.32, 0.… 0.54\*\*\* (0.44, 0.… 0.54\*\*\* (0.43, 0.…  
## 10 wealth\_ind… 0.47\*\*\* (0.38, 0.… 0.65\*\*\* (0.51, 0.… 0.65\*\*\* (0.51, 0.…  
## # ℹ 18 more rows

## Models Validation

### Accessing Multicollinearity (VIF)

# VIFs check (A VIF value > 5 indicates high multicollinearity)  
# Base model  
model\_1\_vif <- vif(baseline\_model)  
print(model\_1\_vif)

## GVIF Df GVIF^(1/(2\*Df))  
## area 1.116394 1 1.056596  
## education\_level 1.673060 5 1.052813  
## wealth\_index 1.418920 4 1.044708  
## health\_insurance 1.018544 1 1.009229  
## current\_contraceptive\_use 1.015178 1 1.007560  
## awareness\_hiv\_aids 1.482419 1 1.217546  
## access\_to\_media 1.263256 1 1.123947

# Enhanced model  
model\_2\_vif <- vif(enhanced\_model)  
print(model\_2\_vif)

## GVIF Df GVIF^(1/(2\*Df))  
## area 1.250833 1 1.118406  
## region 4.773168 5 1.169178  
## education\_level 1.957540 5 1.069476  
## ethnicity 4.140370 3 1.267185  
## wealth\_index 1.953237 4 1.087287  
## health\_insurance 1.049221 1 1.024315  
## current\_contraceptive\_use 1.022997 1 1.011433  
## awareness\_hiv\_aids 1.649486 1 1.284323  
## access\_to\_media 1.310264 1 1.144668

# Interacton model  
model\_3\_vif <- vif(model\_interaction)

## there are higher-order terms (interactions) in this model  
## consider setting type = 'predictor'; see ?vif

print(model\_3\_vif)

## GVIF Df GVIF^(1/(2\*Df))  
## area 12.769190 1 3.573400  
## region 4.828080 5 1.170516  
## education\_level 4107.374849 5 2.298034  
## ethnicity 4.236501 3 1.272041  
## wealth\_index 2.014490 4 1.091492  
## health\_insurance 1.052067 1 1.025703  
## current\_contraceptive\_use 1.023631 1 1.011747  
## awareness\_hiv\_aids 1.650085 1 1.284556  
## access\_to\_media 1.309351 1 1.144269  
## area:education\_level 5655.260451 5 2.372713

* Model 1 has the least concern with multicollinearity.
* Model 2 introduces region and ethnicity, with a mild increase in multicollinearity, but not at alarming levels.
* Model 3 exhibits more noticeable multicollinearity, particularly with the area variable and its interaction with education\_level; their VIF values are significantly higher (around 3.57, 2.3, and 2.37, respectively). Although these values are below 5, the increase is substantial compared to the previous models and could start to affect the reliability and interpretability of the regression coefficients for these variables.
* High multicollinearity in models, especially those with interaction terms, doesn’t invalidate the model but does complicate the interpretation of specific coefficients.

### Likelihood Ratio Test (LRT)

# Model 1 vs. Model 2  
anova(baseline\_model, enhanced\_model, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: child\_marriage ~ area + education\_level + wealth\_index + health\_insurance +   
## current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media  
## Model 2: child\_marriage ~ area + region + education\_level + ethnicity +   
## wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 9253 8000.9   
## 2 9245 7749.7 8 251.12 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The Model 2 provides a significantly better fit to the data compared to the Model 1, as indicated by the large decrease in residual deviance and the very small p-value (< 2.2e-16).

# Model 1 vs. Model 3  
anova(baseline\_model, model\_interaction, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: child\_marriage ~ area + education\_level + wealth\_index + health\_insurance +   
## current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media  
## Model 2: child\_marriage ~ area + region + education\_level + ethnicity +   
## wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media + area:education\_level  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 9253 8000.9   
## 2 9240 7743.7 13 257.14 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The Model 3 provides a significantly better fit to the data compared to the Model 1, as indicated by the large decrease in residual deviance and the very small p-value (< 2.2e-16).

# Model 2 vs. Model 3  
anova(enhanced\_model, model\_interaction, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: child\_marriage ~ area + region + education\_level + ethnicity +   
## wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media  
## Model 2: child\_marriage ~ area + region + education\_level + ethnicity +   
## wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media + area:education\_level  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 9245 7749.7   
## 2 9240 7743.7 5 6.0212 0.3042

The p-value is not significant.

### Assessing AIC and BIC

# Create a data frame to hold AIC and BIC values  
aic\_bic\_comparison <- data.frame(  
 Model = c("Model 1", "Model 2", "Model 3"),  
 AIC = c(AIC(baseline\_model), AIC(enhanced\_model), AIC(model\_interaction)),  
 BIC = c(BIC(baseline\_model), BIC(enhanced\_model), BIC(model\_interaction))  
)  
  
# Print the table  
print(aic\_bic\_comparison)

## Model AIC BIC  
## 1 Model 1 8030.853 8137.868  
## 2 Model 2 7795.737 7959.826  
## 3 Model 3 7799.716 7999.477

Model 2 has the lowest AIC and BIC values of all, indicating that it might be the best model among the three in terms of balancing fit and complexity.

### Hosmer-Lemeshow Test

# 1. Hosmer-Lemeshow Test for the Model 1  
hoslem.test(baseline\_model$y, fitted(baseline\_model), g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: baseline\_model$y, fitted(baseline\_model)  
## X-squared = 12.943, df = 8, p-value = 0.1138

# 2. Hosmer-Lemeshow Test for the Model 2 (Baseline + Fixed Effects)  
hoslem.test(enhanced\_model$y, fitted(enhanced\_model), g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: enhanced\_model$y, fitted(enhanced\_model)  
## X-squared = 5.3687, df = 8, p-value = 0.7175

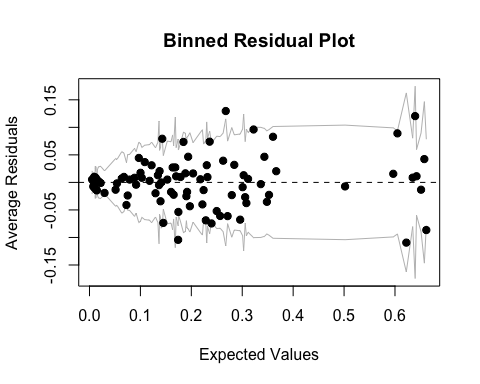
# 3. Hosmer-Lemeshow Test for the Model 3 (Baseline + Fixed Effects + Interaction Terms)  
hoslem.test(model\_interaction$y, fitted(model\_interaction), g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: model\_interaction$y, fitted(model\_interaction)  
## X-squared = 6.1988, df = 8, p-value = 0.625

A large p-value (>0.05) indicates a good fit, meaning that there’s no significant difference between the observed and predicted values. Through each model, the p-value increases which suggests that our decision to include fixed effects and interaction terms are significant.

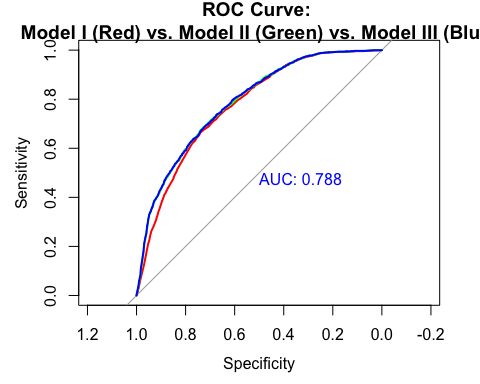
### Residuals Plot

binnedplot(fitted(model\_interaction),  
 residuals(model\_interaction, type = "response"),  
 nclass = NULL,  
 xlab = "Expected Values",  
 ylab = "Average Residuals",  
 main = "Binned Residual Plot",  
 cex.pts = 1,  
 col.int = "gray")



### ROC Curve and AUC

invisible(plot(roc(imputed\_df$child\_marriage,  
 fitted(baseline\_model)),  
 col = "red",  
 main = "ROC Curve: \nModel I (Red) vs. Model II (Green) vs. Model III (Blue)"))  
  
invisible(plot(roc(imputed\_df$child\_marriage,  
 fitted(enhanced\_model)),  
 col = "green",  
 add = T))  
  
invisible(plot(roc(imputed\_df$child\_marriage,  
 fitted(model\_interaction)),  
 print.auc = T,  
 col = "blue",  
 add = T))



# For each model  
roc\_response\_model\_1 <- roc(imputed\_df$child\_marriage, fitted(baseline\_model))  
auc\_model\_1 <- auc(roc\_response\_model\_1)  
  
roc\_response\_model\_2 <- roc(imputed\_df$child\_marriage, fitted(enhanced\_model))  
auc\_model\_2 <- auc(roc\_response\_model\_2)  
  
roc\_response\_model\_3 <- roc(imputed\_df$child\_marriage, fitted(model\_interaction))  
auc\_model\_3 <- auc(roc\_response\_model\_3)  
  
# Compare AUC values  
print(paste("AUC Model 1:", auc\_model\_1))

## [1] "AUC Model 1: 0.772320518006648"

print(paste("AUC Model 2:", auc\_model\_2))

## [1] "AUC Model 2: 0.786801269233356"

print(paste("AUC Model 3:", auc\_model\_3))

## [1] "AUC Model 3: 0.787500065329211"

1. Model 1 (AUC: 0.7723):

An AUC of 0.5 represents a model with no discriminative ability (akin to random guessing), while an AUC of 1.0 represents perfect discrimination. So, my model is performing substantially better than random guessing.

The AUC of 0.7 indicates that the Model 1 has good discriminative ability. In other words, it is capable of distinguishing between cases and controls with a high degree of accuracy.

1. Model 2 (AUC: 0.7868):

This model shows a slight improvement in AUC over the Model 1. The increase suggests that the additional variables (or adjustments) I made in the Model 2 contribute positively to its ability to differentiate between cases and controls. The difference in AUC between the Model 1 and Model 2, while modest, is still meaningful, especially in practical, real-world contexts.

1. Model 3 (AUC: 0.7876):

The Model 3 shows a very slight improvement in AUC over the Model 2. This indicates that adding interaction terms provides a marginal improvement in the model’s discriminatory power. However, the improvement is very minimal, which aligns with my earlier findings that the interaction terms did not significantly improve the model fit.

1. Overall:

All models demonstrate good ability to distinguish between cases and controls. An AUC greater than 0.7 is generally considered acceptable, and my models are above 0.7 and close to 0.8. The Model 2 and Model 3 only show marginal improvements in AUC compared to the Model 1. This suggests that while the additional complexity (more variables and interaction terms) does contribute slightly to model performance, the gains are not substantial.

# 6. Approach 2: Logistic Regression Analysis with No Data Imputation

*After careful consideration regarding data reliability, this project will proceed with Approach 2.*

## Pre-Analysis

### Recreate DataFrames with Removed NAs

# No imputation -- Remove rows with any missing data from 'female\_df'  
og\_df <- na.omit(married\_df)  
print(og\_df)

## # A tibble: 7,591 × 11  
## age\_first\_marriage marital\_status area region education\_level  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 23 1 1 1 5  
## 2 17 1 1 1 2  
## 3 23 1 1 1 5  
## 4 22 1 1 1 5  
## 5 26 1 1 1 5  
## 6 23 1 1 1 5  
## 7 27 1 1 1 4  
## 8 25 1 1 1 5  
## 9 21 1 1 1 3  
## 10 25 1 1 1 5  
## # ℹ 7,581 more rows  
## # ℹ 6 more variables: health\_insurance <dbl>, ethnicity <dbl>,  
## # wealth\_index <dbl>, current\_contraceptive\_use <dbl>,  
## # awareness\_hiv\_aids <dbl>, access\_to\_media <dbl>

### Recreate Necessary Variables

# Convert "age at first marriage" into a binary variable to indicate child marriage  
# Child marriage is defined as marriage before the age of 18  
# The new binary variable "child\_marriage" will have a value of 1 if the marriage occurred before age 18, and 0 otherwise  
og\_df <- og\_df %>%  
 mutate(child\_marriage = ifelse(age\_first\_marriage < 18, 1, 0))

# Create a binary variable for child marriage under 16  
# The new variable "child\_marriage\_u16" will have a value of 1 if the marriage occurred before age 16, and 0 otherwise  
og\_df <- og\_df %>%  
 mutate(child\_marriage\_u16 = ifelse(age\_first\_marriage < 16, 1, 0))

# Move "child\_marriage" and "child\_marriage\_u16" to the front of the dataframe  
og\_df <- og\_df %>%  
 dplyr::select(child\_marriage, child\_marriage\_u16, everything())

### Converting Categorical and Binary Variables to Factors

# Convert nominal and ordinal variables to factors  
og\_df$area <- as.factor(og\_df$area)  
og\_df$region <- as.factor(og\_df$region)  
og\_df$education\_level <- factor(og\_df$education\_level, ordered = FALSE)  
og\_df$ethnicity <- as.factor(og\_df$ethnicity)  
og\_df$wealth\_index <- factor(og\_df$wealth\_index, ordered = FALSE)  
  
# Binary variables are already in the correct format and can be used as is

#write.csv(og\_df, "og\_df.csv", row.names = FALSE)

## Model 1: Base Model

# Base model  
base\_model\_no\_impute <- glm(child\_marriage ~ area + education\_level + wealth\_index + health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media, family = binomial(), data = og\_df)  
  
# Summarize the base model  
summary(base\_model\_no\_impute)

##   
## Call:  
## glm(formula = child\_marriage ~ area + education\_level + wealth\_index +   
## health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids +   
## access\_to\_media, family = binomial(), data = og\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.12285 0.15692 -0.783 0.43370   
## area2 0.26196 0.09295 2.818 0.00483 \*\*   
## education\_level1 -0.43948 0.10272 -4.278 1.88e-05 \*\*\*  
## education\_level2 -0.51284 0.10331 -4.964 6.90e-07 \*\*\*  
## education\_level3 -1.27682 0.13682 -9.332 < 2e-16 \*\*\*  
## education\_level4 -3.38049 0.59104 -5.720 1.07e-08 \*\*\*  
## education\_level5 -3.10583 0.27962 -11.107 < 2e-16 \*\*\*  
## wealth\_index2 -0.61689 0.09276 -6.651 2.92e-11 \*\*\*  
## wealth\_index3 -0.95525 0.11163 -8.558 < 2e-16 \*\*\*  
## wealth\_index4 -0.75531 0.12062 -6.262 3.80e-10 \*\*\*  
## wealth\_index5 -1.01614 0.15960 -6.367 1.93e-10 \*\*\*  
## health\_insurance 0.09667 0.09195 1.051 0.29308   
## current\_contraceptive\_use -0.08254 0.06980 -1.182 0.23704   
## awareness\_hiv\_aids -0.47372 0.08239 -5.750 8.94e-09 \*\*\*  
## access\_to\_media 0.01077 0.10367 0.104 0.91726   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 7404.1 on 7590 degrees of freedom  
## Residual deviance: 6086.8 on 7576 degrees of freedom  
## AIC: 6116.8  
##   
## Number of Fisher Scoring iterations: 7

## Model 2: Base Model with Fixed Effects

# Enhanced model with additional fixed effects  
enhanced\_model\_no\_impute <- glm(child\_marriage ~ area + region + education\_level + ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media, family = binomial(), data = og\_df)  
  
# Summarize the new model with FEs  
summary(enhanced\_model\_no\_impute)

##   
## Call:  
## glm(formula = child\_marriage ~ area + region + education\_level +   
## ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media, family = binomial(),   
## data = og\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.6510890 0.2290549 -7.208 5.67e-13 \*\*\*  
## area2 0.1470706 0.0980382 1.500 0.13358   
## region2 0.0495982 0.1566174 0.317 0.75148   
## region3 -0.0227665 0.1516838 -0.150 0.88069   
## region4 -0.1639096 0.1668163 -0.983 0.32582   
## region5 -0.0308200 0.1398462 -0.220 0.82557   
## region6 0.0503335 0.1461564 0.344 0.73056   
## education\_level1 0.1358036 0.1169557 1.161 0.24558   
## education\_level2 0.0826556 0.1189010 0.695 0.48695   
## education\_level3 -0.7178808 0.1495591 -4.800 1.59e-06 \*\*\*  
## education\_level4 -2.8361281 0.5948134 -4.768 1.86e-06 \*\*\*  
## education\_level5 -2.5784165 0.2856381 -9.027 < 2e-16 \*\*\*  
## ethnicity2 0.5559934 0.1376108 4.040 5.34e-05 \*\*\*  
## ethnicity3 0.3371900 0.1396231 2.415 0.01574 \*   
## ethnicity4 1.9238662 0.1608494 11.961 < 2e-16 \*\*\*  
## wealth\_index2 -0.0532348 0.1077890 -0.494 0.62139   
## wealth\_index3 -0.3666087 0.1293455 -2.834 0.00459 \*\*   
## wealth\_index4 -0.1573256 0.1414044 -1.113 0.26588   
## wealth\_index5 -0.4268584 0.1831853 -2.330 0.01980 \*   
## health\_insurance -0.0958692 0.0954439 -1.004 0.31516   
## current\_contraceptive\_use 0.0008332 0.0728182 0.011 0.99087   
## awareness\_hiv\_aids -0.0914918 0.0944656 -0.969 0.33278   
## access\_to\_media 0.1818904 0.1103596 1.648 0.09932 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 7404.1 on 7590 degrees of freedom  
## Residual deviance: 5849.7 on 7568 degrees of freedom  
## AIC: 5895.7  
##   
## Number of Fisher Scoring iterations: 7

## Model 3: Base Model with Fixed Effects and Interaction Terms

# Adding the interaction term between area and education level  
interaction\_model\_no\_impute <- update(enhanced\_model\_no\_impute, . ~ . + area:education\_level)  
summary(interaction\_model\_no\_impute)

##   
## Call:  
## glm(formula = child\_marriage ~ area + region + education\_level +   
## ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media + area:education\_level,   
## family = binomial(), data = og\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.44930 0.39341 -3.684 0.00023 \*\*\*  
## area2 -0.07873 0.36089 -0.218 0.82730   
## region2 0.04910 0.15677 0.313 0.75412   
## region3 -0.02527 0.15186 -0.166 0.86783   
## region4 -0.15765 0.16694 -0.944 0.34499   
## region5 -0.01975 0.14022 -0.141 0.88798   
## region6 0.05481 0.14624 0.375 0.70781   
## education\_level1 0.13163 0.37724 0.349 0.72714   
## education\_level2 -0.19969 0.36500 -0.547 0.58431   
## education\_level3 -1.04485 0.39533 -2.643 0.00822 \*\*   
## education\_level4 -3.26882 1.06401 -3.072 0.00213 \*\*   
## education\_level5 -2.69970 0.49148 -5.493 3.95e-08 \*\*\*  
## ethnicity2 0.55976 0.13773 4.064 4.82e-05 \*\*\*  
## ethnicity3 0.32597 0.13973 2.333 0.01965 \*   
## ethnicity4 1.94289 0.16204 11.990 < 2e-16 \*\*\*  
## wealth\_index2 -0.05987 0.10795 -0.555 0.57918   
## wealth\_index3 -0.36871 0.12941 -2.849 0.00438 \*\*   
## wealth\_index4 -0.15023 0.14140 -1.062 0.28805   
## wealth\_index5 -0.41743 0.18406 -2.268 0.02334 \*   
## health\_insurance -0.09355 0.09555 -0.979 0.32754   
## current\_contraceptive\_use 0.00196 0.07286 0.027 0.97854   
## awareness\_hiv\_aids -0.09512 0.09437 -1.008 0.31351   
## access\_to\_media 0.17945 0.11022 1.628 0.10349   
## area2:education\_level1 -0.01759 0.39520 -0.045 0.96450   
## area2:education\_level2 0.32521 0.38034 0.855 0.39252   
## area2:education\_level3 0.39737 0.42012 0.946 0.34423   
## area2:education\_level4 0.59025 1.28428 0.460 0.64581   
## area2:education\_level5 0.03787 0.62518 0.061 0.95170   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 7404.1 on 7590 degrees of freedom  
## Residual deviance: 5846.0 on 7563 degrees of freedom  
## AIC: 5902  
##   
## Number of Fisher Scoring iterations: 7

# 7. Approach Comparison:

# Extracting data from the models  
extract\_model\_data <- function(model) {  
 model\_summary <- summary(model)  
 coeffs <- model\_summary$coefficients  
 data.frame(  
 Term = rownames(coeffs),  
 Estimate = sprintf("%.3f", coeffs[, "Estimate"]),  
 pValue = ifelse(coeffs[, "Pr(>|z|)"] < 0.001,   
 format(coeffs[, "Pr(>|z|)"], scientific = TRUE),  
 sprintf("%.3f", coeffs[, "Pr(>|z|)"])),  
 Significance = sapply(coeffs[, "Pr(>|z|)"], add\_asterisks)  
 )  
}

## Model 1: Base Model

# Applying the function to each model  
base\_impute\_data <- extract\_model\_data(baseline\_model)  
base\_no\_impute\_data <- extract\_model\_data(base\_model\_no\_impute)  
  
# Combine the data for comparison  
base\_comparison\_data <- merge(base\_impute\_data, base\_no\_impute\_data, by = "Term", suffixes = c("\_Impute", "\_NoImpute"), sort = FALSE)  
  
# View the comparison  
print(base\_comparison\_data)

## Term Estimate\_Impute pValue\_Impute Significance\_Impute  
## 1 (Intercept) -0.351 0.009 \*\*  
## 2 area2 0.408 5.222876e-07 \*\*\*  
## 3 education\_level1 -0.365 3.696088e-05 \*\*\*  
## 4 education\_level2 -0.434 6.270603e-07 \*\*\*  
## 5 education\_level3 -1.116 1.076550e-21 \*\*\*  
## 6 education\_level4 -3.144 8.474392e-10 \*\*\*  
## 7 education\_level5 -2.929 1.060085e-30 \*\*\*  
## 8 wealth\_index2 -0.550 1.230578e-11 \*\*\*  
## 9 wealth\_index3 -0.934 1.867146e-20 \*\*\*  
## 10 wealth\_index4 -0.761 5.702313e-12 \*\*\*  
## 11 wealth\_index5 -1.021 1.156489e-11 \*\*\*  
## 12 health\_insurance 0.117 0.154   
## 13 current\_contraceptive\_use -0.033 0.603   
## 14 awareness\_hiv\_aids -0.453 1.412117e-10 \*\*\*  
## 15 access\_to\_media -0.027 0.741   
## Estimate\_NoImpute pValue\_NoImpute Significance\_NoImpute  
## 1 -0.123 0.434   
## 2 0.262 0.005 \*\*  
## 3 -0.439 1.881675e-05 \*\*\*  
## 4 -0.513 6.901996e-07 \*\*\*  
## 5 -1.277 1.036155e-20 \*\*\*  
## 6 -3.380 1.067977e-08 \*\*\*  
## 7 -3.106 1.155012e-28 \*\*\*  
## 8 -0.617 2.917388e-11 \*\*\*  
## 9 -0.955 1.153337e-17 \*\*\*  
## 10 -0.755 3.797438e-10 \*\*\*  
## 11 -1.016 1.931436e-10 \*\*\*  
## 12 0.097 0.293   
## 13 -0.083 0.237   
## 14 -0.474 8.940563e-09 \*\*\*  
## 15 0.011 0.917

* Most variables show consistent results in terms of significance and direction of effect across both models. This suggests that the imputation process has not drastically altered the relationships between these predictors and the outcome variable.
* Education and Wealth Index variables are consistent, robust and significant predictors of the outcome across both models.
* Significance of Area: This variable’s significance suggests geographical variation is an important factor, but its influence is reduced in the non-imputed model.
* The overall patterns suggest that while some predictors are consistently influential, others are sensitive to how missing data is handled.

## Model 2: Base Model with Fixed Effects

# Applying the function to each model  
enhanced\_impute\_data <- extract\_model\_data(enhanced\_model)  
enhanced\_no\_impute\_data <- extract\_model\_data(enhanced\_model\_no\_impute)  
  
# Combine the data for comparison  
enhanced\_comparison\_data <- merge(enhanced\_impute\_data, enhanced\_no\_impute\_data, by = "Term", suffixes = c("\_Impute", "\_NoImpute"), sort = FALSE)  
  
# View the comparison  
print(enhanced\_comparison\_data)

## Term Estimate\_Impute pValue\_Impute Significance\_Impute  
## 1 (Intercept) -1.194 6.082446e-11 \*\*\*  
## 2 area2 0.268 0.002 \*\*  
## 3 region2 0.364 0.005 \*\*  
## 4 region3 0.162 0.209   
## 5 region4 0.331 0.010 \*\*  
## 6 region5 -0.009 0.942   
## 7 region6 -0.042 0.757   
## 8 education\_level1 0.023 0.811   
## 9 education\_level2 -0.032 0.742   
## 10 education\_level3 -0.740 2.491054e-09 \*\*\*  
## 11 education\_level4 -2.816 4.474974e-08 \*\*\*  
## 12 education\_level5 -2.598 6.330343e-24 \*\*\*  
## 13 ethnicity2 -0.004 0.968   
## 14 ethnicity3 0.284 0.029 \*  
## 15 ethnicity4 1.200 4.023378e-28 \*\*\*  
## 16 wealth\_index2 -0.231 0.008 \*\*  
## 17 wealth\_index3 -0.612 1.033996e-08 \*\*\*  
## 18 wealth\_index4 -0.433 2.596704e-04 \*\*\*  
## 19 wealth\_index5 -0.685 2.334946e-05 \*\*\*  
## 20 health\_insurance 0.003 0.973   
## 21 current\_contraceptive\_use 0.032 0.623   
## 22 awareness\_hiv\_aids -0.185 0.017 \*  
## 23 access\_to\_media -0.081 0.353   
## Estimate\_NoImpute pValue\_NoImpute Significance\_NoImpute  
## 1 -1.651 5.666765e-13 \*\*\*  
## 2 0.147 0.134   
## 3 0.050 0.751   
## 4 -0.023 0.881   
## 5 -0.164 0.326   
## 6 -0.031 0.826   
## 7 0.050 0.731   
## 8 0.136 0.246   
## 9 0.083 0.487   
## 10 -0.718 1.586814e-06 \*\*\*  
## 11 -2.836 1.859743e-06 \*\*\*  
## 12 -2.578 1.766623e-19 \*\*\*  
## 13 0.556 5.337557e-05 \*\*\*  
## 14 0.337 0.016 \*  
## 15 1.924 5.710335e-33 \*\*\*  
## 16 -0.053 0.621   
## 17 -0.367 0.005 \*\*  
## 18 -0.157 0.266   
## 19 -0.427 0.020 \*  
## 20 -0.096 0.315   
## 21 0.001 0.991   
## 22 -0.091 0.333   
## 23 0.182 0.099

* Consistent findings across the imputed and non-imputed models in certain variables (e.g., education levels 3, 4, 5, and ethnicity4) suggest that imputation does not radically alter these relationships. This consistency can indicate that imputation is not introducing significant bias for these variables.
* Imputation might reveal statistically significant relationships that are not evident in the non-imputed data (e.g., area2, region2, region4, awareness of HIV/AIDS).
* For some variables (e.g., wealth\_index categories), the significance and estimate values change between the models.
* The loss of significance in many variables in the non-imputed model suggests that the missing data may not be random and could be related to the variables’ influence on the outcome.

## Model 3: Base Model with Fixed Effects and Interaction Terms

# Applying the function to each model  
interaction\_impute\_data <- extract\_model\_data(model\_interaction)  
interaction\_no\_impute\_data <- extract\_model\_data(interaction\_model\_no\_impute)  
  
# Combine the data for comparison  
interaction\_comparison\_data <- merge(interaction\_impute\_data, interaction\_no\_impute\_data, by = "Term", suffixes = c("\_Impute", "\_NoImpute"), sort = FALSE)  
  
# View the comparison  
print(interaction\_comparison\_data)

## Term Estimate\_Impute pValue\_Impute Significance\_Impute  
## 1 (Intercept) -0.805 0.008 \*\*  
## 2 area2 -0.149 0.589   
## 3 region2 0.353 0.007 \*\*  
## 4 region3 0.156 0.228   
## 5 region4 0.331 0.010 \*\*  
## 6 region5 0.003 0.980   
## 7 region6 -0.042 0.754   
## 8 education\_level1 -0.154 0.612   
## 9 education\_level2 -0.519 0.067   
## 10 education\_level3 -1.184 2.034807e-04 \*\*\*  
## 11 education\_level4 -3.614 5.074377e-04 \*\*\*  
## 12 education\_level5 -2.979 8.897857e-12 \*\*\*  
## 13 ethnicity2 -0.003 0.977   
## 14 ethnicity3 0.268 0.039 \*  
## 15 ethnicity4 1.222 1.268874e-28 \*\*\*  
## 16 wealth\_index2 -0.244 0.005 \*\*  
## 17 wealth\_index3 -0.621 7.055082e-09 \*\*\*  
## 18 wealth\_index4 -0.432 2.777919e-04 \*\*\*  
## 19 wealth\_index5 -0.680 3.219988e-05 \*\*\*  
## 20 health\_insurance 0.009 0.912   
## 21 current\_contraceptive\_use 0.035 0.588   
## 22 awareness\_hiv\_aids -0.192 0.013 \*  
## 23 access\_to\_media -0.085 0.329   
## 24 area2:education\_level1 0.176 0.580   
## 25 area2:education\_level2 0.547 0.064   
## 26 area2:education\_level3 0.500 0.140   
## 27 area2:education\_level4 1.024 0.391   
## 28 area2:education\_level5 0.410 0.449   
## Estimate\_NoImpute pValue\_NoImpute Significance\_NoImpute  
## 1 -1.449 2.296503e-04 \*\*\*  
## 2 -0.079 0.827   
## 3 0.049 0.754   
## 4 -0.025 0.868   
## 5 -0.158 0.345   
## 6 -0.020 0.888   
## 7 0.055 0.708   
## 8 0.132 0.727   
## 9 -0.200 0.584   
## 10 -1.045 0.008 \*\*  
## 11 -3.269 0.002 \*\*  
## 12 -2.700 3.951423e-08 \*\*\*  
## 13 0.560 4.817267e-05 \*\*\*  
## 14 0.326 0.020 \*  
## 15 1.943 4.015586e-33 \*\*\*  
## 16 -0.060 0.579   
## 17 -0.369 0.004 \*\*  
## 18 -0.150 0.288   
## 19 -0.417 0.023 \*  
## 20 -0.094 0.328   
## 21 0.002 0.979   
## 22 -0.095 0.314   
## 23 0.179 0.103   
## 24 -0.018 0.965   
## 25 0.325 0.393   
## 26 0.397 0.344   
## 27 0.590 0.646   
## 28 0.038 0.952

* Consistency in the significance of variables like education levels 3, 4, and 5, and ethnicity4 across both models suggests that imputation is not drastically changing these relationships.
* The imputed model shows significant effects in variables (like region2, region4, ethnicity3, and interaction terms) that are not significant in the non-imputed model.
* The differing levels of significance for variables like wealth index categories and awareness of HIV/AIDS between models raise questions about how missing data impacts these variables.
* The loss of significance in many variables in the non-imputed model suggests that the missing data may not be random and could be related to the variables’ influence on the outcome.

# 8. Models Comparison (Odd Ratios and 95% Confidence Intervals)

# Tidy the baseline model with confidence intervals  
tidy\_base\_no\_impute <- tidy(base\_model\_no\_impute, conf.int = TRUE, exponentiate = TRUE)  
  
# Tidy the enhanced model with confidence intervals  
tidy\_enhanced\_no\_impute <- tidy(enhanced\_model\_no\_impute, conf.int = TRUE, exponentiate = TRUE)  
  
# Tidy the interaction model with confidence intervals  
tidy\_interaction\_no\_impute <- tidy(interaction\_model\_no\_impute, conf.int = TRUE, exponentiate = TRUE)

# Apply the function to each model's p.value  
tidy\_base\_no\_impute$asterisks <- sapply(tidy\_base\_no\_impute$p.value, add\_asterisks)  
tidy\_enhanced\_no\_impute$asterisks <- sapply(tidy\_enhanced\_no\_impute$p.value, add\_asterisks)  
tidy\_interaction\_no\_impute$asterisks <- sapply(tidy\_interaction\_no\_impute$p.value, add\_asterisks)

# Create OR strings with asterisks and format CIs as a string  
tidy\_base\_no\_impute <- tidy\_base\_no\_impute %>%  
 mutate(  
 OR = ifelse(is.na(estimate), NA, paste0(round(estimate, 2), asterisks)),  
 CI = ifelse(is.na(conf.low) | is.na(conf.high), NA, format\_ci(conf.low, conf.high))  
 )  
  
tidy\_enhanced\_no\_impute <- tidy\_enhanced\_no\_impute %>%  
 mutate(  
 OR = ifelse(is.na(estimate), NA, paste0(round(estimate, 2), asterisks)),  
 CI = ifelse(is.na(conf.low) | is.na(conf.high), NA, format\_ci(conf.low, conf.high))  
 )  
  
tidy\_interaction\_no\_impute <- tidy\_interaction\_no\_impute %>%  
 mutate(  
 OR = ifelse(is.na(estimate), NA, paste0(round(estimate, 2), asterisks)),  
 CI = ifelse(is.na(conf.low) | is.na(conf.high), NA, format\_ci(conf.low, conf.high))  
 )

# Add a 'Model' column to each tidied dataframe  
tidy\_base\_no\_impute <- tidy\_base\_no\_impute %>% mutate(Model = "Model\_1")  
tidy\_enhanced\_no\_impute <- tidy\_enhanced\_no\_impute %>% mutate(Model = "Model\_2")  
tidy\_interaction\_no\_impute <- tidy\_interaction\_no\_impute %>% mutate(Model = "Model\_3")  
  
# Combine and pivot the dataframes  
combined\_no\_impute <- bind\_rows(  
 tidy\_base\_no\_impute %>% dplyr::select(term, OR, CI, Model),  
 tidy\_enhanced\_no\_impute %>% dplyr::select(term, OR, CI, Model),  
 tidy\_interaction\_no\_impute %>% dplyr::select(term, OR, CI, Model)  
) %>%  
 pivot\_wider(names\_from = Model, values\_from = c(OR, CI))

# Replace NAs with "—"  
combined\_no\_impute[is.na(combined\_no\_impute)] <- "—"

# Reordering columns to have OR and CI next to each other for each model  
combined\_no\_impute <- combined\_no\_impute %>%  
 dplyr::select(term,   
 OR\_Model\_1, CI\_Model\_1,   
 OR\_Model\_2, CI\_Model\_2,   
 OR\_Model\_3, CI\_Model\_3)

# Print the final combined table  
print(combined\_no\_impute)

## # A tibble: 28 × 7  
## term OR\_Model\_1 CI\_Model\_1 OR\_Model\_2 CI\_Model\_2 OR\_Model\_3 CI\_Model\_3  
## <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 (Intercept) 0.88 (0.65, 1.… 0.19\*\*\* (0.12, 0.… 0.23\*\*\* (0.1, 0.4…  
## 2 area2 1.3\*\* (1.08, 1.… 1.16 (0.96, 1.… 0.92 (0.47, 1.…  
## 3 education\_… 0.64\*\*\* (0.53, 0.… 1.15 (0.91, 1.… 1.14 (0.56, 2.…  
## 4 education\_… 0.6\*\*\* (0.49, 0.… 1.09 (0.86, 1.… 0.82 (0.41, 1.…  
## 5 education\_… 0.28\*\*\* (0.21, 0.… 0.49\*\*\* (0.36, 0.… 0.35\*\* (0.17, 0.…  
## 6 education\_… 0.03\*\*\* (0.01, 0.… 0.06\*\*\* (0.01, 0.… 0.04\*\* (0, 0.21)   
## 7 education\_… 0.04\*\*\* (0.02, 0.… 0.08\*\*\* (0.04, 0.… 0.07\*\*\* (0.03, 0.…  
## 8 wealth\_ind… 0.54\*\*\* (0.45, 0.… 0.95 (0.77, 1.… 0.94 (0.76, 1.…  
## 9 wealth\_ind… 0.38\*\*\* (0.31, 0.… 0.69\*\* (0.54, 0.… 0.69\*\* (0.54, 0.…  
## 10 wealth\_ind… 0.47\*\*\* (0.37, 0.… 0.85 (0.65, 1.… 0.86 (0.65, 1.…  
## # ℹ 18 more rows

1. (Intercept) Odds Ratios and Significance:

* In Model 1, the odds ratio for the intercept is 0.88, which indicates a slight negative association, but it’s not statistically significant, as indicated by the absence of asterisks.
* In Model 2, the intercept has an OR of 0.19, which is highly significant (indicated by three asterisks) and suggests a much stronger negative association.
* Model 3 shows a similar pattern to Model 2, with a significant OR of 0.27, indicating a robust negative association.

1. Area2:

* This term has an OR of 1.3 in Model 1 and is statistically significant (denoted by two asterisks), suggesting a moderately positive association with the outcome.
* In Models 2 and 3, the ORs are 1.16 and 1.15, respectively, showing a slight positive association, but these are not statistically significant.

1. Education Level Categories:

* All education levels in Model 1 show a significant negative association (ORs less than 1, with three asterisks), meaning higher education levels are associated with lower odds of the outcome.
* In Models 2 and 3, education\_level1 and education\_level2 lose their statistical significance (ORs 1.15 and 1.09 in Model 2, and 1.14 and 1.08 in Model 3), indicating no strong association.
* However, education\_level3, education\_level4, and education\_level5 maintain their strong, significant negative association in Models 2 and 3, as indicated by the ORs (0.49, 0.06, and 0.08) and the presence of three asterisks.

1. Wealth Index Categories:

* In Model 1, all wealth index categories show a significant negative association with the outcome, with ORs ranging from 0.36 to 0.54 and marked with three asterisks.
* The strength of this association decreases in Models 2 and 3, with ORs moving closer to 1 and varying levels of significance, showing that wealth’s impact on the outcome is less pronounced or potentially confounded by other variables in these models.

1. Health Insurance:

* There is no statistically significant association between health insurance and the outcome in any model, as indicated by ORs around 1 and the absence of asterisks.

1. Current Contraceptive Use:

* Similar to health insurance, this variable shows no significant association with the outcome in any of the models.

1. Awareness of HIV/AIDS:

* In Model 1, there is a significant negative association (OR 0.62, three asterisks), but this association becomes non-significant in Models 2 and 3.

1. Access to Media:

This variable shows no consistent or significant association with the outcome across the three models.

1. Region Variables in Models 2 and 3:

* These variables are only included in the latter two models and show varying ORs, but none of them are statistically significant.

1. Ethnicity Categories in Models 2 and 3:

* Ethnicity shows significant positive associations in these models, particularly for ethnicity4, which has very high ORs and is highly significant.

1. Interaction Terms ethnicity:access\_to\_media in Model 3:

* The interaction terms involving ethnicity and access to media are introduced in Model 3, but none show statistical significance.

# 9. Models Validation

## Accessing Multicollinearity (VIF)

# VIFs check (A VIF value > 5 indicates high multicollinearity)  
# Base model  
model\_1\_vif <- vif(base\_model\_no\_impute)  
print(model\_1\_vif)

## GVIF Df GVIF^(1/(2\*Df))  
## area 1.178945 1 1.085793  
## education\_level 1.745961 5 1.057313  
## wealth\_index 1.613551 4 1.061629  
## health\_insurance 1.029280 1 1.014534  
## current\_contraceptive\_use 1.009914 1 1.004945  
## awareness\_hiv\_aids 1.483532 1 1.218003  
## access\_to\_media 1.216401 1 1.102906

# Enhanced model  
model\_2\_vif <- vif(enhanced\_model\_no\_impute)  
print(model\_2\_vif)

## GVIF Df GVIF^(1/(2\*Df))  
## area 1.311056 1 1.145014  
## region 5.810682 5 1.192402  
## education\_level 2.259331 5 1.084921  
## ethnicity 7.849140 3 1.409733  
## wealth\_index 2.856573 4 1.140199  
## health\_insurance 1.087155 1 1.042667  
## current\_contraceptive\_use 1.025894 1 1.012864  
## awareness\_hiv\_aids 1.784748 1 1.335945  
## access\_to\_media 1.260862 1 1.122881

# Interacton model  
model\_3\_vif <- vif(interaction\_model\_no\_impute)

## there are higher-order terms (interactions) in this model  
## consider setting type = 'predictor'; see ?vif

print(model\_3\_vif)

## GVIF Df GVIF^(1/(2\*Df))  
## area 17.655196 1 4.201809  
## region 5.841703 5 1.193037  
## education\_level 2372.303552 5 2.175289  
## ethnicity 8.072108 3 1.416330  
## wealth\_index 2.921655 4 1.143414  
## health\_insurance 1.088687 1 1.043402  
## current\_contraceptive\_use 1.026229 1 1.013030  
## awareness\_hiv\_aids 1.783196 1 1.335364  
## access\_to\_media 1.259876 1 1.122442  
## area:education\_level 3313.050828 5 2.249173

* Model 1 has the least concern with multicollinearity.
* Model 2 introduces region and ethnicity, with a mild increase in multicollinearity, but not at alarming levels.
* Model 3 exhibits more noticeable multicollinearity, particularly with area, education\_level and its interaction with one another`; their VIF values are significantly higher (around 4.2, 2.17, and 2.25, respectively). Although these values are below 5, the increase is substantial compared to the previous models and could start to affect the reliability and interpretability of the regression coefficients for these variables.
* High multicollinearity in models, especially those with interaction terms, doesn’t invalidate the model but does complicate the interpretation of specific coefficients.

## Likelihood Ratio Test (LRT)

# Model 1 vs. Model 2  
anova(base\_model\_no\_impute, enhanced\_model\_no\_impute, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: child\_marriage ~ area + education\_level + wealth\_index + health\_insurance +   
## current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media  
## Model 2: child\_marriage ~ area + region + education\_level + ethnicity +   
## wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 7576 6086.8   
## 2 7568 5849.7 8 237.05 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The Model 2 provides a significantly better fit to the data compared to the Model 1, as indicated by the large decrease in residual deviance and the very small p-value (< 2.2e-16).

# Model 1 vs. Model 3  
anova(base\_model\_no\_impute, interaction\_model\_no\_impute, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: child\_marriage ~ area + education\_level + wealth\_index + health\_insurance +   
## current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media  
## Model 2: child\_marriage ~ area + region + education\_level + ethnicity +   
## wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media + area:education\_level  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 7576 6086.8   
## 2 7563 5846.0 13 240.78 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The Model 3 provides a significantly better fit to the data compared to the Model 1, as indicated by the large decrease in residual deviance and the very small p-value (< 2.2e-16).

# Model 2 vs. Model 3  
anova(enhanced\_model\_no\_impute, interaction\_model\_no\_impute, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: child\_marriage ~ area + region + education\_level + ethnicity +   
## wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media  
## Model 2: child\_marriage ~ area + region + education\_level + ethnicity +   
## wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media + area:education\_level  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 7568 5849.7   
## 2 7563 5846.0 5 3.7254 0.5896

The residual deviance in Model 3 is slightly lower compared to Model 1, indicating a potential improvement in model fit but not significant.

The p-value of 0.5896 is not below the common significance level. This indicates that the addition of the interaction term area:education\_level in Model 3 does not significantly improve the model fit compared to Model 2. In simpler terms, the interaction between area and education\_level does not seem to have a significant effect on predicting child marriage.

## Assessing AIC and BIC

# Create a data frame to hold AIC and BIC values  
aic\_bic\_comparison <- data.frame(  
 Model = c("Model 1", "Model 2", "Model 3"),  
 AIC = c(AIC(base\_model\_no\_impute), AIC(enhanced\_model\_no\_impute), AIC(interaction\_model\_no\_impute)),  
 BIC = c(BIC(base\_model\_no\_impute), BIC(enhanced\_model\_no\_impute), BIC(interaction\_model\_no\_impute))  
)  
  
# Print the table  
print(aic\_bic\_comparison)

## Model AIC BIC  
## 1 Model 1 6116.798 6220.819  
## 2 Model 2 5895.745 6055.243  
## 3 Model 3 5902.019 6096.191

Model 2 has the lowest AIC and BIC values of all, indicating that it might be the best model among the three in terms of balancing fit and complexity.

## Hosmer-Lemeshow Test

# 1. Hosmer-Lemeshow Test for the Model 1  
hoslem.test(base\_model\_no\_impute$y, fitted(base\_model\_no\_impute), g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: base\_model\_no\_impute$y, fitted(base\_model\_no\_impute)  
## X-squared = 15.295, df = 8, p-value = 0.05366

# 2. Hosmer-Lemeshow Test for the Model 2 (Baseline + Fixed Effects)  
hoslem.test(enhanced\_model\_no\_impute$y, fitted(enhanced\_model\_no\_impute), g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: enhanced\_model\_no\_impute$y, fitted(enhanced\_model\_no\_impute)  
## X-squared = 10.076, df = 8, p-value = 0.2597

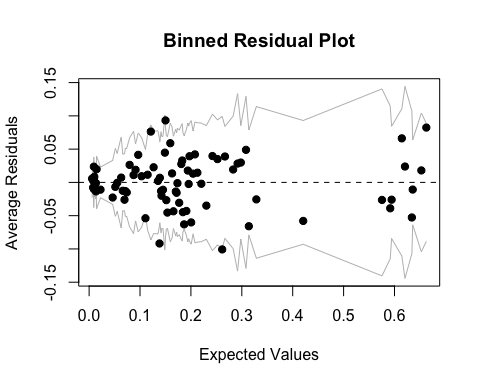
# 3. Hosmer-Lemeshow Test for the Model 3 (Baseline + Fixed Effects + Interaction Terms)  
hoslem.test(interaction\_model\_no\_impute$y, fitted(interaction\_model\_no\_impute), g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: interaction\_model\_no\_impute$y, fitted(interaction\_model\_no\_impute)  
## X-squared = 6.9744, df = 8, p-value = 0.5394

A large p-value (>0.05) indicates a good fit, meaning that there’s no significant difference between the observed and predicted values. Through each model, the p-value increases which suggests that our decision to include fixed effects and interaction terms are significant.

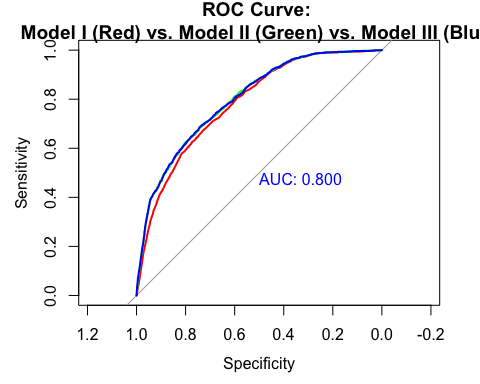
## Residuals Plot

binnedplot(fitted(interaction\_model\_no\_impute),  
 residuals(interaction\_model\_no\_impute, type = "response"),  
 nclass = NULL,  
 xlab = "Expected Values",  
 ylab = "Average Residuals",  
 main = "Binned Residual Plot",  
 cex.pts = 1,  
 col.int = "gray")



## ROC Curve and AUC

invisible(plot(roc(og\_df$child\_marriage,  
 fitted(base\_model\_no\_impute)),  
 col = "red",  
 main = "ROC Curve: \nModel I (Red) vs. Model II (Green) vs. Model III (Blue)"))  
  
invisible(plot(roc(og\_df$child\_marriage,  
 fitted(enhanced\_model\_no\_impute)),  
 col = "green",  
 add = T))  
  
invisible(plot(roc(og\_df$child\_marriage,  
 fitted(interaction\_model\_no\_impute)),  
 print.auc = T,  
 col = "blue",  
 add = T))



# For each model  
roc\_model\_1 <- roc(og\_df$child\_marriage, fitted(base\_model\_no\_impute))  
auc\_mod\_1 <- auc(roc\_model\_1)  
  
roc\_model\_2 <- roc(og\_df$child\_marriage, fitted(enhanced\_model\_no\_impute))  
auc\_mod\_2 <- auc(roc\_model\_2)  
  
roc\_model\_3 <- roc(og\_df$child\_marriage, fitted(interaction\_model\_no\_impute))  
auc\_mod\_3 <- auc(roc\_model\_3)  
  
# Compare AUC values  
print(paste("AUC Model 1:", auc\_mod\_1))

## [1] "AUC Model 1: 0.78337409946712"

print(paste("AUC Model 2:", auc\_mod\_2))

## [1] "AUC Model 2: 0.79976685814396"

print(paste("AUC Model 3:", auc\_mod\_3))

## [1] "AUC Model 3: 0.800031108041481"

1. Model 1 (AUC: 0.7834):

An AUC of 0.5 represents a model with no discriminative ability (akin to random guessing), while an AUC of 1.0 represents perfect discrimination. So, my model is performing substantially better than random guessing.

The AUC of 0.7 indicates that the Model 1 has good discriminative ability. In other words, it is capable of distinguishing between cases and controls with a high degree of accuracy.

1. Model 2 (AUC: 0.7998):

This model shows a slight improvement in AUC over the Model 1. The increase suggests that the additional variables (or adjustments) I made in the Model 2 contribute positively to its ability to differentiate between cases and controls. The difference in AUC between the Model 1 and Model 2, while modest, is still meaningful, especially in practical, real-world contexts.

1. Model 3 (AUC: 0.8000):

The Model 3 shows a very slight improvement in AUC over the Model 2. This indicates that adding interaction terms provides a marginal improvement in the model’s discriminatory power. However, the improvement is very minimal, which aligns with my earlier findings that the interaction terms did not significantly improve the model fit.

1. Overall:

All models demonstrate good ability to distinguish between cases and controls. An AUC greater than 0.7 is generally considered acceptable, and my models are above 0.7 and model 3 equals to 0.8. The Model 2 and Model 3 only show marginal improvements in AUC compared to the Model 1. This suggests that while the additional complexity (more variables and interaction terms) does contribute slightly to model performance, the gains are not substantial.