PHC 6089 - R Project

Marina Alacoque Rodrigues

2025-10-01

1 Executive Summary

The use of substances such as tobacco and alcohol during pregnancy poses a serious risk to both maternal and fetal health. Although the dangers are well-known, a portion of pregnant women continue to use these substances, influenced by a complex set of social, demographic, and mental health factors. The primary objective of this project is to assess the relationship between these factors and substance use in a nationally representative sample of pregnant women in the United States. Using data from the National Survey on Drug Use and Health (NSDUH), this study will employ logistic regression models to identify the most significant predictors of substance use during pregnancy. The results can inform public health policies and intervention strategies targeted at the most vulnerable subgroups.

2 Background

2.1 National Survey on Drug Use and Health

This study used data from the National Survey on Drug Use and Health (NSDUH), a nationally representative, cross-sectional household survey conducted annually among non-institutionalized U.S. residents. The NSDUH aimed to monitor patterns of drug and alcohol use, mental health status, and associated health behaviors. Its complex sampling design, featuring clustered and stratified sampling and oversampling of specific subgroups (such as young individuals, Hispanic, and Black/African-American populations), ensured precise estimates while maintaining representativeness of the U.S. population.

The sampling weights in the survey were assigned to each respondent to account for the unequal probability of selection, nonresponse, and post-stratification adjustments aligned with U.S. Census population estimates. This approach allowed the generation of design-based estimates, such as means, proportions, and standard errors that accurately reflected population parameters. In addition, the survey included a wide range of demographic, behavioral, and clinical variables, enabling the analysis of complex associations between substance use and health outcomes. All statistical analyses were therefore performed considering the survey's design variables, weights, strata, and primary sampling units (PSUs), to produce unbiased and valid national estimates.

Given this methodological rigor, the NSDUH represents a powerful data source to investigate substance use during pregnancy, an important public health concern associated with premature birth, low birth weight, and congenital syndromes. Identifying both modifiable and non-modifiable risk factors is essential for the development of effective prevention and cessation programs. By leveraging this robust and recent national database, the present study provided an updated overview of substance use among pregnant women in the U.S., offering valuable evidence to support healthcare professionals in screening and counseling at-risk populations.

2.2 Objects

The objectives of this study were:

1. To characterize the sociodemographic and mental health profile of pregnant women in the NSDUH sample.
2. To estimate the prevalence of tobacco and alcohol use among pregnant women in the sample.
3. To evaluate the association between sociodemographic factors (age, race/ethnicity, education, income), marital status, and mental health status with substance use during pregnancy.

3 Methods

3.1 Study design and data

A cross-sectional study was conducted using the most recent publicly available data (2023) from the National Survey on Drug Use and Health (NSDUH). The dataset is managed and distributed by the Substance Abuse and Mental Health Data Archive (SAMHDA), under the sponsorship of the Substance Abuse and Mental Health Services Administration (SAMHSA) of the U.S. Department of Health and Human Services.

The final sample included women aged 12 to 49 years who participated in the 2023 NSDUH and reported being pregnant at the time of the interview. Participants with missing data for the primary variables of interest were excluded from the multivariate analyses to ensure the validity of the results.

The 2023 NSDUH dataset consists of a tibble data frame (tbl_df) containing 56,705 observations and 2,636 variables. Each row represents an individual respondent, while the columns capture a comprehensive range of sociodemographic, behavioral, and mental health indicators. The tibble structure facilitates efficient data handling, manipulation, and analysis, allowing for robust exploration of complex associations despite the large number of variables.

3.2 Study variables

Information were extracted from the general questionnaires and specific modules of the NSDUH.

Outcome Variables: Binary variables (Yes/No) were created for the use of each substance of interest in the last month and last year: Use of tobacco (cigarettes). Use of alcohol.

Predictor Variables (Exposures): Sociodemographic: Age group, Race/Ethnicity, Education level, Annual household income, Residence (urban/rural). Clinical and Behavioral: Mental health status (e.g., presence of a Major Depressive Episode in the past year, persistent stress), Marital Status.

3.3 Statistical Analysis

All analysis were performed using R software version 4.4.3.

1. Descriptive Analysis: The characteristics of the sample of pregnant women were summarized using frequencies and proportions for categorical variables and means with standard deviations for continuous numerical variables.

2. Bivariate Analysis: The initial association between each predictor variable and the outcomes (tobacco/alcohol use) was assessed with the chi-square test of independence.

3. Multivariate Analysis: To evaluate the adjusted association, logistic regression models were constructed for each outcome. The results were presented as adjusted Odds Ratios (aOR) with their respective 95% confidence intervals. The analysis allowed for the examination of the relationship between risk factors and substance use while simultaneously controlling for potential confounding variables.

3.4 Data Preparation

The data were provided in Stata (.dta) file format and imported into R for analysis. Initially, the required R packages were installed and loaded to enable data management and design-based statistical analysis.

```r
library(survey)
```

```
## Warning: pacote 'survey' foi compilado no R versão 4.4.3

## Carregando pacotes exigidos: grid

## Carregando pacotes exigidos: Matrix

## Carregando pacotes exigidos: survival

## Warning: pacote 'survival' foi compilado no R versão 4.4.3

##
## Anexando pacote: 'survey'

## O seguinte objeto é mascarado por 'package:graphics':
##
##     dotchart
```

```r
library(svydiags)
```

```
## Warning: pacote 'svydiags' foi compilado no R versão 4.4.3

## Carregando pacotes exigidos: MASS
```

```r
library(tidyverse)
```

```
## Warning: pacote 'tidyverse' foi compilado no R versão 4.4.3

## Warning: pacote 'ggplot2' foi compilado no R versão 4.4.3

## Warning: pacote 'tibble' foi compilado no R versão 4.4.3

## Warning: pacote 'tidyr' foi compilado no R versão 4.4.3

## Warning: pacote 'readr' foi compilado no R versão 4.4.3

## Warning: pacote 'purrr' foi compilado no R versão 4.4.3

## Warning: pacote 'dplyr' foi compilado no R versão 4.4.3

## Warning: pacote 'stringr' foi compilado no R versão 4.4.3

## Warning: pacote 'forcats' foi compilado no R versão 4.4.3

## Warning: pacote 'lubridate' foi compilado no R versão 4.4.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x dplyr::select() masks MASS::select()
## x tidyr::unpack() masks Matrix::unpack()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(naniar)
```

```
## Warning: pacote 'naniar' foi compilado no R versão 4.4.3
```

```r
library(stats)
library(jtools)
```

```
## Warning: pacote 'jtools' foi compilado no R versão 4.4.3
```

```r
library(pander)
```

```
## Warning: pacote 'pander' foi compilado no R versão 4.4.3
```

```r
library(dplyr)
library(sjlabelled)
```

```
## Warning: pacote 'sjlabelled' foi compilado no R versão 4.4.3
```

```
## 
## Anexando pacote: 'sjlabelled'
## 
## O seguinte objeto é mascarado por 'package:forcats':
## 
##     as_factor
## 
## O seguinte objeto é mascarado por 'package:dplyr':
## 
##     as_label
## 
## O seguinte objeto é mascarado por 'package:ggplot2':
## 
##     as_label
```

```r
library(broom)
```

```
## Warning: pacote 'broom' foi compilado no R versão 4.4.3
```

```r
library(gridExtra)
```

```
## Warning: pacote 'gridExtra' foi compilado no R versão 4.4.3
```

```
## 
## Anexando pacote: 'gridExtra'
## 
## O seguinte objeto é mascarado por 'package:dplyr':
## 
##     combine
```

```r
library(haven)
```

```
## Warning: pacote 'haven' foi compilado no R versão 4.4.3
```

```
## 
## Anexando pacote: 'haven'
## 
## Os seguintes objetos são mascarados por 'package:sjlabelled':
## 
##     as_factor, read_sas, read_spss, read_stata, write_sas, zap_labels
```

```r
library(dplyr)

dados <- read_dta("C:/Users/55319/OneDrive/PHC6089/R project/data/NSDUH_2023.dta")
head(dados)
```

```
## # A tibble: 6 x 2,636
##   QUESTID2 FILEDATE   ANALWT2_C VESTR_C VEREP PDEN10   COUTYP4 MAIIN102 AIIND102
##      <dbl> <chr>          <dbl>   <dbl> <dbl> <dbl+lb> <dbl+l> <dbl+lb> <dbl+lb>
## 1 10000053 03/25/2025     3276.   40031     2 2 [2 - ~ 2 [2 -~ 2 [2 - ~ 2 [2 - ~
## 2 10000679 03/25/2025    15630.   40021     2 2 [2 - ~ 3 [3 -~ 2 [2 - ~ 2 [2 - ~
## 3 10001208 03/25/2025     4018.   40043     1 2 [2 - ~ 2 [2 -~ 2 [2 - ~ 2 [2 - ~
## 4 10001260 03/25/2025    10712.   40030     2 2 [2 - ~ 2 [2 -~ 2 [2 - ~ 2 [2 - ~
## 5 10001588 03/25/2025     8195.   40023     2 2 [2 - ~ 2 [2 -~ 2 [2 - ~ 2 [2 - ~
## 6 10004996 03/25/2025     3771.   40048     1 2 [2 - ~ 2 [2 -~ 2 [2 - ~ 2 [2 - ~
## # i 2,627 more variables: AGE3 <dbl+lbl>, SERVICE <dbl+lbl>, MILSTAT <dbl+lbl>,
## #   ACTDEVER <dbl+lbl>, ACTD2001 <dbl+lbl>, ACTD9001 <dbl+lbl>,
## #   ACTD7590 <dbl+lbl>, ACTDVIET <dbl+lbl>, ACTDPRIV <dbl+lbl>,
## #   COMBATPY <dbl+lbl>, NOMARR2 <dbl+lbl>, HEALTH <dbl+lbl>,
## #   MOVSINPYR2 <dbl+lbl>, SEXATRACT2 <dbl+lbl>, SEXIDENT22 <dbl+lbl>,
## #   SPEAKENGL <dbl+lbl>, LVLDIFSEE2 <dbl+lbl>, LVLDIFHEAR2 <dbl+lbl>,
## #   LVLDIFWALK2 <dbl+lbl>, LVLDIFMEM2 <dbl+lbl>, LVLDIFCARE2 <dbl+lbl>, ...
```

```r
class(dados)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```r
dim(dados)
```

```
## [1] 56705  2636
```

In this study, the subpopulation of interest were female adults (aged 12 to 49 years old), in the United States, who reported being pregnant at the time of the interview, with non-missing data on the following variables: past-year tobacco use (TOBYR), past-month tobacco use (TOBMON), past-year alcohol use (ALCYR), past-month alcohol use (ALCMON), any health insurance (ANYHLTI2), age category (CATAG6), race/ethnicity (NEWRACE2), education level (EDUHIGHCAT), annual household income (INCOME), marital status (IRMARIT), past-year major depressive episode (AMDEYR), and psychological distress (KSSLR6MONED). The variables that were irrelevant for this analysis purposes were removed from the subset.

Following the procedures below, we identified a subpopulation of N = 598 participants who had met our study eligibility criteria.

```r
nsduh <- dados

# Setting invalid responses to missing (NA)
nsduh <- nsduh %>%
  replace_with_na(replace = list(
    SEXIDENT = c(85, 94, 97, 98, 99),
    ANYHLTI2 = c(94, 97, 98, 99),
    PREG = c(98, 99),
    TOBYR = c(98, 99),
    TOBMON = c(98, 99),
    ALCYR = c(98, 99),
    ALCMON = c(98, 99),
    EDUHIGHCAT = c(99),
    IRMARIT = c(99),
    INCOME = c(99)
  ))
```

```
## Warning: Missing from data: `SEXIDENT`
```

```
# Creating factor variables for the original categorical variables
categorical_vars <- c("PREG", "IRSEX", "AMDEYR", "NEWRACE2", "CATAG6", "INCOME",
                      "ANYHLTI2", "EDUHIGHCAT", "IRMARIT", "TOBYR", "TOBMON",
                      "ALCYR", "ALCMON")
nsduh <- nsduh %>%
  mutate(across(all_of(categorical_vars), as.factor))

# Creating clean, labeled variables for analysis
nsduh <- nsduh %>%
  mutate(
    woman = factor(ifelse(IRSEX == 2, "Yes", "No")),
    depression_bin = factor(ifelse(AMDEYR == 1, "Yes", "No")),
    healthins = factor(ifelse(ANYHLTI2 == 1, "Yes", "No")),
    stress_cont = KSSLR6MONED,
    age_cat = factor(CATAG6, labels = c("12-17", "18-25", "26-34", "35-49",
                                        "50-64", "65+")),
    race_eth = factor(NEWRACE2, labels = c("White", "Black", "Nat Am/AK Nat",
                                           "Hawaiian/PI", "Asian",
                                           "Multiracial", "Hispanic")),
    education = factor(EDUHIGHCAT, labels = c("< High School",
                                              "High School Grad",
                                              "Some College",
                                              "College Grad", "Other")),
    education = fct_collapse(education,"Some College" = c("Some College",
                                                          "Other")),
    marital_status = factor(IRMARIT, labels = c("Married",
                                                "Widowed/Divorced/Separated",
                                                "Never Married",
                                                "Living with partner")),
    income_level = factor(INCOME, labels = c("< $20k", "$20k-$49,999", "$50k-$74,999",
                                             ">= $75k")),
    tobacco_yr = factor(ifelse(TOBYR == 1, "Yes", "No")),
    tobacco_mon = factor(ifelse(TOBMON == 1, "Yes", "No")),
    alcohol_yr = factor(ifelse(ALCYR == 1, "Yes", "No")),
    alcohol_mon = factor(ifelse(ALCMON == 1, "Yes", "No"))
  )

# Creating the 'subpop' variable to identify complete cases for the analysis
nsduh <- nsduh %>%
  mutate(
    subpop = factor(ifelse(
      !is.na(tobacco_yr) & !is.na(alcohol_yr) & !is.na(age_cat) &
      !is.na(race_eth) & !is.na(education) & !is.na(income_level) &
      !is.na(marital_status) & !is.na(healthins) & !is.na(depression_bin) &
      !is.na(stress_cont),
      1, 0
    ))
  )

# Verifying the count of complete cases
```

```
table(nsduh$subpop)
```

```
##
##     0     1
## 14077 42628
```

```
# Filtering for the study population (pregnant women with complete data)
final_subset_preg <- nsduh %>%
  filter(PREG == 1 & subpop == 1) %>%
  dplyr::select(
    QUESTID2, tobacco_yr, tobacco_mon, alcohol_yr, alcohol_mon,
    age_cat, race_eth, education, income_level, marital_status,
    healthins, depression_bin, stress_cont,
    ANALWT2_C, VESTR_C, VEREP
  )

# Checking the dimensions of the final dataframe
dim(final_subset_preg)
```

```
## [1] 598  16
```

```
# (A) Adjusting the person-level weight
# Since we are using data from only one wave (2023),
# no weight adjustment (division by number of waves) is needed.

weight_adj <- final_subset_preg$ANALWT2_C

# (B) Creating survey design object
# This step defines the complex survey design to ensure that all analyses
# (means, proportions, regressions, etc.) account for the stratified and clustered
# sampling structure.

nsduh_design <- svydesign(
  id = ~VEREP,            # Primary sampling units (clusters)
  strata = ~VESTR_C,      # Variance estimation strata
  weights = ~weight_adj,  # Adjusted person-level weight
  data = final_subset_preg,
  nest = TRUE
)
```

4 Results

4.1 Summary of Social and Demographic Variables

Tables 1 and 2 provide a summary of key demographic, behavioral, and health characteristics of pregnant women included in the NSDUH 2023 dataset. Table 1 presents descriptive statistics for numeric variables, including age and stress scores, allowing assessment of central tendency and variability within the population. Table 2 summarizes categorical variables, such as race/ethnicity, education, marital status, health insurance coverage, depression status, and substance use, reporting counts and proportions for each category. Together, these tables offer a comprehensive overview of the sample, supporting the contextual interpretation of subsequent analyses.

```
# Creating numeric age variable from categories

final_subset_preg <- final_subset_preg %>%
  mutate(
    age_cont = case_when(
      age_cat == "12-17" ~ 14.5,
      age_cat == "18-25" ~ 21.5,
      age_cat == "26-34" ~ 30,
      age_cat == "35-49" ~ 42,
      age_cat == "50-64" ~ 57,
      age_cat == "65+" ~ 70,
      TRUE ~ NA_real_
    )
  )

# Adding descriptive labels to variables

final_subset_preg <- final_subset_preg %>%
  mutate(
    age_cont = set_label(age_cont, "Age (years)"),
    stress_cont = set_label(stress_cont, "Stress Score"),
    race_eth = set_label(race_eth, "Race/Ethnicity"),
    education = set_label(education, "Education Level"),
    income_level = set_label(income_level,"Income Level"),
    marital_status = set_label(marital_status, "Marital Status"),
    healthins = set_label(healthins, "Has Health Insurance"),
    depression_bin = set_label(depression_bin, "Depression in Past Year"),
    tobacco_mon = set_label(tobacco_mon, "Tobacco Use Past Month"),
    alcohol_mon = set_label(alcohol_mon, "Alcohol Use Past Month")
  )

nsduh_design$variables <- final_subset_preg

# Numeric summary table (Age and Stress)

numeric_vars <- c("age_cont", "stress_cont")

numeric_summary_df <- lapply(numeric_vars, function(var){
  mean_val <- coef(svymean(as.formula(paste0("~", var)), nsduh_design, na.rm=TRUE))[1]
  se_val   <- SE(svymean(as.formula(paste0("~", var)), nsduh_design, na.rm=TRUE))[1]
  n_val    <- sum(weights(nsduh_design))  # weighted N
  min_val  <- svyquantile(as.formula(paste0("~", var)), nsduh_design, 0, na.rm=TRUE,
                          method="constant")[[1]]
  max_val  <- svyquantile(as.formula(paste0("~", var)), nsduh_design, 1, na.rm=TRUE,
                          method="constant")[[1]]

  data.frame(
    Variable = get_label(final_subset_preg[[var]]),
    N = round(n_val),       # weighted N
    Mean = round(mean_val,2),
    SE = round(se_val,2),
    Min = round(min_val,2),
    Max = round(max_val,2),
    stringsAsFactors = FALSE
  )
}) %>% bind_rows()
```

```r
numeric_summary_df_clean <- as.data.frame(numeric_summary_df, stringsAsFactors = FALSE)
rownames(numeric_summary_df_clean) <- NULL

numeric_summary_df_clean <- numeric_summary_df_clean[, c("Variable", "N", "Mean",
                                                         "SE", "Min.quantile",
                                                         "Max.quantile")]
cols_to_format_num <- c("N", "Mean", "SE", "Min.quantile", "Max.quantile")

numeric_summary_df_display <- numeric_summary_df_clean %>%
  mutate(across(all_of(cols_to_format_num),
                ~ ifelse(. == 0, "~ 0", as.character(.))))

pander(numeric_summary_df_display,
       caption = "Weighted numeric summaries of age and stress scores among
       pregnant women, NSDUH 2023.",
       missing = "")
```

Table 1: Weighted numeric summaries of age and stress scores among pregnant women, NSDUH 2023.

Variable	N	Mean	SE	Min.quantile	Max.quantile
Age (years)	1876615	30.52	0.63	21.5	42
Stress Score	1876615	4.3	0.29	~ 0	24

```r
visual_table <- function(vars, design, data) {
  result <- list()

  for (var in vars) {
    var_label <- get_label(data[[var]])
    if (is.null(var_label) || length(var_label) == 0) {
      var_label <- var
    }
    tab <- svytable(as.formula(paste0("~", var)), design)
    prop_tab <- prop.table(tab)

    # Add variable header row
    result <- append(result, list(data.frame(
      Variable = get_label(data[[var]]),
      Category = "",
      N = NA,
      Percent = NA
    )))

    # Add category rows
    cat_df <- data.frame(
      Variable = "",
      Category = names(tab),
      N = round(as.numeric(tab), 0),
      Percent = round(as.numeric(prop_tab),3),
      stringsAsFactors = FALSE
    )
    result <- append(result, list(cat_df))
  }

  final_df <- bind_rows(result)
  return(final_df)
}

#  Categorical variables

categorical_vars <- c("race_eth", "education", "marital_status", "income_level",
                      "healthins", "depression_bin", "tobacco_mon", "alcohol_mon")

# Categorical summary table

categorical_summary_df <- visual_table(categorical_vars, nsduh_design, final_subset_preg)
tab_cross <- svytable(~ tobacco_mon + alcohol_mon, nsduh_design)
prop_cross <- prop.table(tab_cross)
prop_both <- prop_cross["Yes", "Yes"]
n_both <- as.numeric(tab_cross["Yes", "Yes"])
both_row <- data.frame(
  Variable = "Tobacco & Alcohol Use Past Month",
  Category = "Yes (both)",
  N = round(n_both, 0),
  Percent = round(prop_both, 3),
  stringsAsFactors = FALSE
)

categorical_summary_df <- bind_rows(categorical_summary_df, both_row)

cols_to_format_cat <- c("N", "Percent")

categorical_summary_df_display <- categorical_summary_df %>%
  mutate(across(all_of(cols_to_format_cat),
                ~ ifelse(. == 0, "~ 0", as.character(.))))

# Displaying categorical table
pander(categorical_summary_df_display,
       justify = c("left","left","right","right"),
       caption = "Weighted categorical summaries of demographic, health,
       and behavioral variables among pregnant women, NSDUH 2023.",  split.table = Inf)
```

Table 2: Weighted categorical summaries of demographic, health, and behavioral variables among pregnant women, NSDUH 2023.

Variable	Category	N	Percent
Race/Ethnicity		NA	NA
	White	1079833	0.575
	Black	304069	0.162
	Nat Am/AK Nat	6535	0.003
	Hawaiian/PI	587	~ 0
	Asian	124825	0.067
	Multiracial	25828	0.014
	Hispanic	334938	0.178

Variable	Category	N	Percent
Education Level		NA	NA
	< High School	84012	0.045
	High School Grad	400283	0.213
	Some College	469913	0.25
	College Grad	922407	0.492
Marital Status		NA	NA
	Married	1203055	0.641
	Widowed/Divorced/Separated	711	~ 0
	Never Married	84448	0.045
	Living with partner	588400	0.314
Income Level		NA	NA
	< $20k	285653	0.152
	$20k-$49,999	421212	0.224
	$50k-$74,999	208109	0.111
	>= $75k	961640	0.512
Has Health Insurance		NA	NA
	No	111831	0.06
	Yes	1764785	0.94
Depression in Past Year		NA	NA
	No	1713636	0.913
	Yes	162980	0.087
Tobacco Use Past Month		NA	NA
	No	1785554	0.951
	Yes	91061	0.049
Alcohol Use Past Month		NA	NA
	No	1725858	0.92
	Yes	150758	0.08
Tobacco & Alcohol Use Past Month	Yes (both)	13217	0.007

The mean age of pregnant women was approximately 30.5 years, ranging from 21.5 to 42 years. The average stress score was 4.3 (SE = 0.29), indicating a moderate level of stress in this population.

Table 2 shows weighted categorical distributions of social, health, and behavioral variables. Most participants identified as White (57.5%), followed by Black (16.2%) and Hispanic (17.8%), with smaller proportions reporting Asian, Multiracial, or Native American/Alaska Native race/ethnicity. Regarding education, nearly half of the sample were college graduates (49.2%), with 25% having some college and 21.3% completing high school. The majority were married (64.1%), while 31.4% reported living with a partner, and only 4.5% had never married. Most participants reported having health insurance (94%), and 8.7% reported experiencing depression in the past year. Tobacco use in the past month was reported by 4.9% of women, whereas alcohol use was more prevalent, reported by 8.0% of participants. Approximately 0.7% of pregnant women reported using both tobacco and alcohol in the past month.

4.2 Evaluation of the relationship between the socio-demographic variables, mental health and tobacco/alcohol use

To investigate Aim 3, the univariate relationship between the sociodemographic variables, mental health and tobacco/alcohol use was examined using chi-squared analyses. The results are presented in Table 3.

```
final_subset_preg <- droplevels(final_subset_preg)

# Predictors
predictors <- c("age_cat", "race_eth", "education","income_level",
```

```r
                "marital_status", "depression_bin", "healthins")

# Outcomes
outcomes <- c("tobacco_mon", "alcohol_mon")


# Creating empty data frame to store results

results <- data.frame(
  Outcome = character(),
  Predictor = character(),
  Chi_square = numeric(),
  df = numeric(),
  p_value = numeric(),
  stringsAsFactors = FALSE
)

nsduh_design_preg <- svydesign(
  id = ~VEREP,
  strata = ~VESTR_C,
  weights = ~weight_adj,
  data = final_subset_preg,
  nest = TRUE
)

# Loop to run tests and save results
for (outcome in outcomes) {
  for (var in predictors) {
    svytest <- svychisq(as.formula(paste0("~", outcome, "+", var)),
                          nsduh_design_preg, statistic = "F")
    test <- data.frame(
        Outcome = outcome,
        Predictor = var,
        Chi_square = as.numeric(svytest$statistic)[1],
        df = as.numeric(svytest$parameter)[1],
        p_value = as.numeric(svytest$p.value)[1],
        stringsAsFactors = FALSE
      )
    results <- bind_rows(results, test)
  }
}
print(nrow(results))
```

```
## [1] 14
```

```r
# Filtering only meaningful associations

results_sig <- results %>%
  filter(!is.na(p_value) & p_value < 0.05) %>%
  arrange(Outcome, Predictor)

# Table with pander
pander(results_sig,
       justify = c("left","left","right","right", "right"),
       caption = "Significant bivariate associations between predictors
       and substance use (survey-weighted Chi-square).",  split.table = Inf)
```

Table 3: Significant bivariate associations between predictors and substance use (survey-weighted Chi-square).

Outcome	Predictor	Chi_square	df	p_value
alcohol_mon	marital_status	4.59	1.745	0.01639
tobacco_mon	depression_bin	6.683	1	0.0127
tobacco_mon	education	10.6	2.852	3.797e-06
tobacco_mon	income_level	13.32	2.637	4.546e-07
tobacco_mon	marital_status	10.17	2.158	5.748e-05

Significant association was identified between marital status and alcohol use ($\chi^2 = 4.59$, p = 0.016), indicating that patterns of alcohol consumption varied by marital status among pregnant women.Regarding tobacco use, significant associations were observed with education level ($\chi^2 = 10.6$, p $<$ 0.001), income level ($\chi^2 = 13.32$, p $<$ 0.001), marital status ($\chi^2 = 10.17$, p $<$ 0.001), and depressive symptoms ($\chi^2 = 6.68$, p = 0.013). These findings suggest that tobacco use during pregnancy is more strongly related to socioeconomic and psychological factors than alcohol use, highlighting the potential influence of educational attainment, economic conditions, and mental health status on smoking behaviors among pregnant women.

Figure 1 provides a visual representation of these associations, displaying the survey-weighted prevalence of past-month tobacco and alcohol use stratified by significant sociodemographic factors.

```
library(ggplot2)
library(scales)    # For formatting Y-axis as proportions
```

```
## Warning: pacote 'scales' foi compilado no R versão 4.4.3

##
## Anexando pacote: 'scales'

## O seguinte objeto é mascarado por 'package:purrr':
##
##     discard

## O seguinte objeto é mascarado por 'package:readr':
##
##     col_factor
```

```
# --- Plot 1: Education vs. Tobacco ---

# Calculating the weighted prevalences
edu_tob_weighted <- svyby(
  ~tobacco_mon,
  by = ~education,
  design = nsduh_design,
  FUN = svymean,
  na.rm = TRUE
)

# Creating the ggplot object (saved as 'p1')
# Note: 'tobacco_monYes' is the prevalence of "Yes",
# and 'se.tobacco_monYes' is the standard error
p1 <- ggplot(edu_tob_weighted, aes(x = education, y = tobacco_monYes,
                                   fill = education)) +
  geom_col() +
  geom_errorbar(
        aes(ymin = pmax(0, tobacco_monYes - se.tobacco_monYes),
        ymax = tobacco_monYes + se.tobacco_monYes),
    width = 0.2
  ) +
  scale_y_continuous(labels = percent) +
  labs(
    title = "Tobacco Use by Education Level",
    y = "Prevalence (Yes)",
    x = "Education"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_blank(), legend.position = "right",
        plot.title = element_text(size = 8),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8),
    legend.title = element_text(size = 8),
    legend.text = element_text(size = 8)  )


# All the steps shown in Plot 1 should be applied to each variable with a
# significant association:


# --- Plot 2: Income Level vs. Tobacco ---

inc_tob_weighted <- svyby(
  ~tobacco_mon,
  by = ~income_level,
  design = nsduh_design,
  FUN = svymean,
  na.rm = TRUE
)

p2 <- ggplot(inc_tob_weighted, aes(x = income_level, y = tobacco_monYes,
                                   fill = income_level)) +
  geom_col() +
  geom_errorbar(
    aes(ymin = pmax(0, tobacco_monYes - se.tobacco_monYes),
        ymax = tobacco_monYes + se.tobacco_monYes),
    width = 0.2
  ) +
  scale_y_continuous(labels = percent) +
  labs(
    title = "Tobacco Use by Income Level",
    y = "Prevalence (Yes)",
    x = "Income Level"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_blank(), legend.position = "right",
        plot.title = element_text(size = 8),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8),
    legend.title = element_text(size = 8),
    legend.text = element_text(size = 8)  )


# --- Plot 3: Marital Status vs. Tobacco ---

mar_tob_weighted <- svyby(
  ~tobacco_mon,
  by = ~marital_status,
  design = nsduh_design,
  FUN = svymean,
  na.rm = TRUE
)

p3 <- ggplot(mar_tob_weighted, aes(x = marital_status, y = tobacco_monYes,
                                   fill = marital_status)) +
  geom_col() +
  geom_errorbar(
    aes(ymin = pmax(0, tobacco_monYes - se.tobacco_monYes),
        ymax = tobacco_monYes + se.tobacco_monYes),
    width = 0.2) +
  scale_y_continuous(labels = percent) +
  labs(title = "Tobacco Use by Marital Status",
    y = "Prevalence (Yes)",
    x = "Marital Status"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_blank(), legend.position = "right",
        plot.title = element_text(size = 8),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8),
    legend.title = element_text(size = 8),
    legend.text = element_text(size = 8)  )


# --- Plot 4: Marital Status vs. Alcohol ---

mar_alc_weighted <- svyby(
  ~alcohol_mon,
  by = ~marital_status,
  design = nsduh_design,
  FUN = svymean,
  na.rm = TRUE
)

p4 <- ggplot(mar_alc_weighted, aes(x = marital_status, y = alcohol_monYes,
                                   fill = marital_status)) +
  geom_col() +
  geom_errorbar(
    aes(ymin = pmax(0, alcohol_monYes - se.alcohol_monYes),
        ymax = alcohol_monYes + se.alcohol_monYes),
    width = 0.2
  ) +
  scale_y_continuous(labels = percent) +
  labs(
    title = "Alcohol Use by Marital Status",
    y = "Prevalence (Yes)",
    x = "Marital Status",
  ) +
  theme_minimal() +
  theme(axis.text.x = element_blank(), legend.position = "right",
        plot.title = element_text(size = 8),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8),
    legend.title = element_text(size = 8),
    legend.text = element_text(size = 8)
  )


# --- Arranging all plots in a 2x2 grid ---

grid.arrange(
  p1, p2, p3, p4,
  ncol = 2,
  nrow = 2,
  top = textGrob("Figure 1: Weighted Prevalence of Substance Use by
                 Sociodemographic Factors",
    gp = gpar(fontsize = 10)
  )
)
```

Figure 1: Weighted Prevalence of Substance Use by Sociodemographic Factors

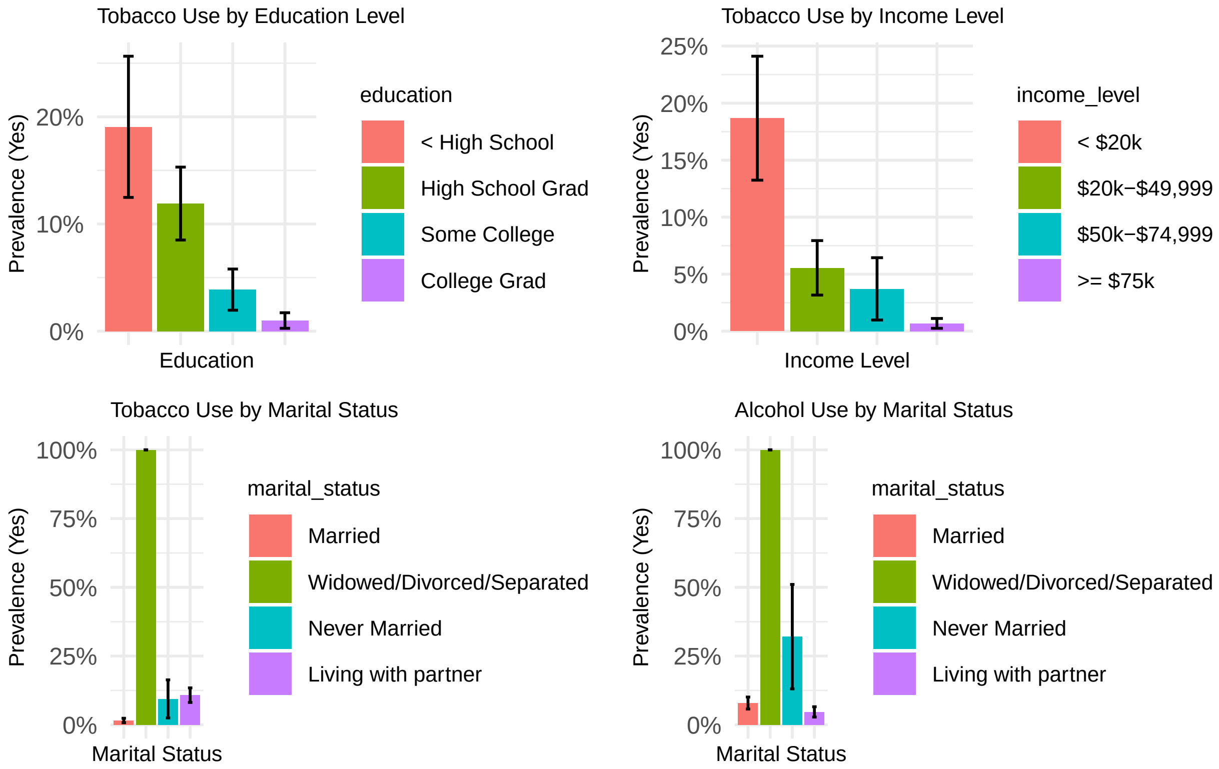


Figure 2 shows the distribution of stress scores by substance use. The boxplots illustrate that participants who reported using tobacco or alcohol in the past month tend to have higher stress scores compared with non-users, suggesting a potential association between elevated stress levels and substance use.

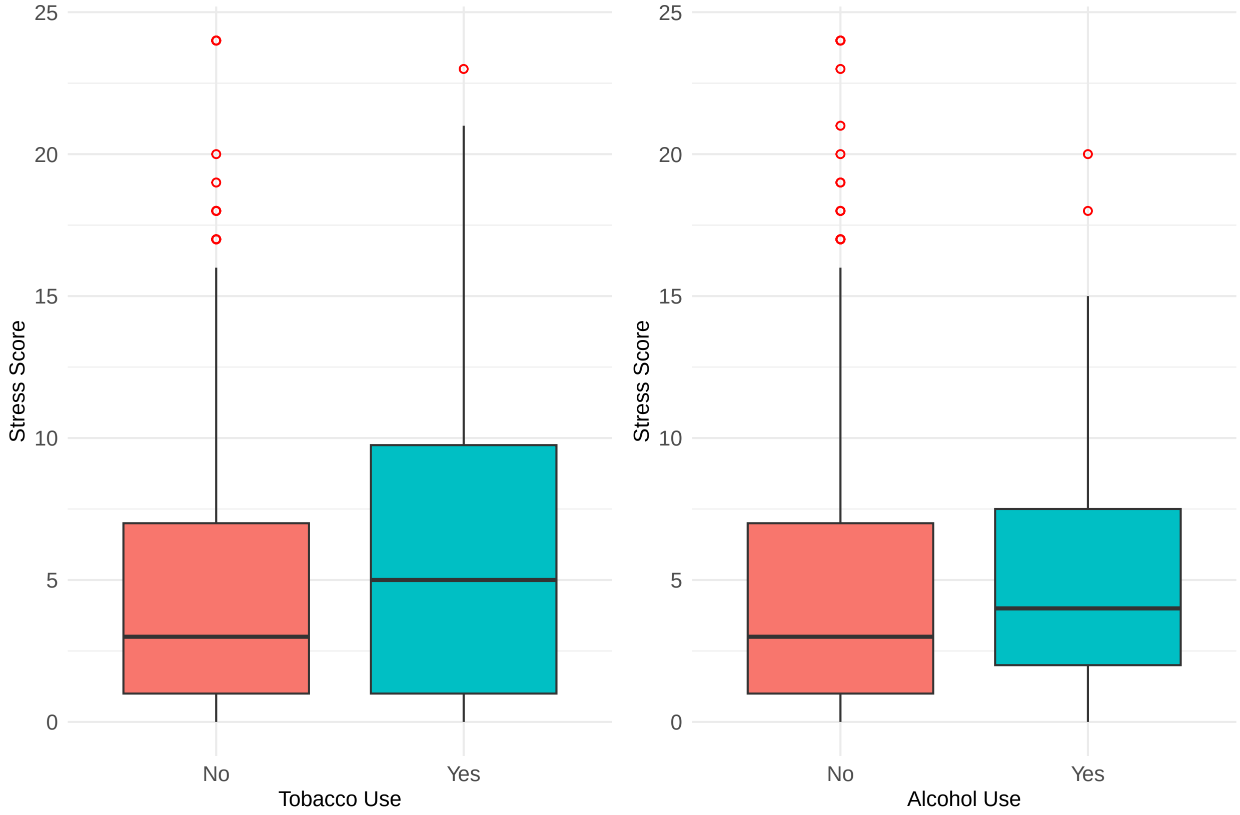
```
library(ggplot2)
library(gridExtra)
library(grid)

# --- Boxplot: Stress vs. Tobacco ---
p_box_tob <- ggplot(nsduh_design$variables, aes(x = tobacco_mon, y = stress_cont, fill = tobacco_mon))
  geom_boxplot(outlier.color = "red", outlier.shape = 1) +
  scale_y_continuous(name = "Stress Score") +
  labs(x = "Tobacco Use"
  ) +
  theme_minimal(base_size = 8) +
  theme(
    legend.position = "none",
    axis.text.x = element_text(size = 8),
    axis.text.y = element_text(size = 8),
    plot.title = element_text(size = 8)
  )

# --- Boxplot: Stress vs. Alcohol ---
p_box_alc <- ggplot(nsduh_design$variables, aes(x = alcohol_mon, y = stress_cont, fill = alcohol_mon))
  geom_boxplot(outlier.color = "red", outlier.shape = 1) +
  scale_y_continuous(name = "Stress Score") +
  labs( x = "Alcohol Use"
  ) +
  theme_minimal(base_size = 8) +
  theme(
    legend.position = "none",
    axis.text.x = element_text(size = 8),
    axis.text.y = element_text(size = 8),
    plot.title = element_text(size = 8)
  )

# --- Arranging both boxplots side by side ---
grid.arrange(
  p_box_tob, p_box_alc,
  ncol = 2,
  top = textGrob("Figure 2. Distribution of Stress Score by Substance Use",
                 gp = gpar(fontsize = 10))
)
```

Figure 2. Distribution of Stress Score by Substance Use



While the study initially aimed to model past-month tobacco use (tobacco_mon), the low prevalence of this outcome (4.9%) in the final analytic sample (N=598) led to severe model instability (sparse data bias) during multivariate logistic regression. To obtain stable and interpretable estimates, the outcome was broadened to past-year tobacco use (tobacco_yr). This variable offered a higher number of events, allowing for a robust and valid regression analysis.Logistic regression models were fitted to examine factors associated with past-year tobacco and alcohol use. Each model was adjusted for education, income level, marital status, stress and depression.

```
library(forcats)

# Cleaning and recoding the variables:

final_subset_preg_clean <- final_subset_preg %>%
  mutate(
    # 1. Recoding Marital Status (2 groups)
    marital_recoded = fct_collapse(marital_status,
      "Married" = "Married",
      "Not Married" = c("Never Married", "Widowed/Divorced/Separated",
                        "Living with partner")
    ),

    # 2. Recoding Education (group "Other" with a valid category)
    education_recoded = fct_collapse(education,
      "< High School" = "< High School",
      "High School Grad" = "High School Grad",
      "Some College" = c("Some College", "Other"),
      "College Grad" = "College Grad"
    ),

    # 3. Recoding Income Level (3 groups for stability)
    income_recoded = fct_collapse(income_level,
      "< $50k" = c("< $20k", "$20k-$49,999"),
      "$50k-$74,999" = "$50k-$74,999",
      ">= $75k" = ">= $75k"
    ), stress_cont = as.numeric(stress_cont)
  ) %>%
     # setting references:
  mutate(
    marital_recoded = relevel(marital_recoded, ref = "Married"),
    education_recoded = relevel(education_recoded, ref = "College Grad"),
    income_recoded = relevel(income_recoded, ref = ">= $75k"),
    depression_bin = relevel(depression_bin, ref = "No"),
   # Reference = No depression
  )
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `education_recoded = fct_collapse(...)`.
## Caused by warning:
## ! Unknown levels in `f`: Other
```

```
# Updating:
nsduh_design_clean <- svydesign(
  id = ~VEREP,
  strata = ~VESTR_C,
  weights = ~weight_adj,
  data = final_subset_preg_clean,
  nest = TRUE
)

# Logistic regression model:
model_tobacco_yr_stable <- svyglm(
  tobacco_yr ~ education_recoded + income_recoded + marital_recoded +
    depression_bin + stress_cont,
  design = nsduh_design_clean,
  family = quasibinomial()
)


results_table <- tidy(model_tobacco_yr_stable, conf.int = TRUE,
                      exponentiate = TRUE) %>%
 dplyr::select(term, estimate, conf.low, conf.high, p.value) %>%
 rename(
    Variable = term,
    `Odds Ratio (OR)` = estimate,
    `95% CI Low` = conf.low,
    `95% CI High` = conf.high,
    `p-value` = p.value
  ) %>%
 mutate(
    Variable = recode(Variable,
      "(Intercept)" = "Intercept",
      "education_recoded< High School" = "Education (Ref: College Grad)
      / < High School",
      "education_recodedHigh School Grad" = "Education (Ref: College Grad)
      / High School Grad",
      "education_recodedSome College" = "Education (Ref: College Grad)
      / Some College",
      "income_recoded< $50k" = "Income (Ref: >= $75k) / < $50k",
      "income_recoded$50k-$74,999" = "Income (Ref: >= $75k) / $50k-$74,999",
      "marital_recodedNot Married" = "Marital Status (Ref: Married)
      / Not Married",
      "depression_binYes" = "Depression (Ref: No) / Yes",
      "stress_cont" = "Stress Score"
    )
  )

pander(
  results_table,
  caption = "Table 4 (Revised): Logistic Regression for Tobacco Use in the Past Year",
  split.table = Inf
)
```

Table 4: Table 4 (Revised): Logistic Regression for Tobacco Use in the Past Year

Variable	Odds Ratio (OR)	95% CI Low	95% CI High	p-value
Intercept	0.01669	0.006939	0.04014	6.62e-12
Education (Ref: College Grad) / < High School	2.303	0.6336	8.369	0.1992
Education (Ref: College Grad) / High School Grad	1.683	0.3681	7.692	0.4934
Education (Ref: College Grad) / Some College	1.278	0.4738	3.449	0.6202
Income (Ref: >= $75k) / < $50k	1.936	0.8136	4.606	0.1316

Variable	Odds Ratio (OR)	95% CI Low	95% CI High	p-value
Income (Ref: >= $75k) / $50k–$74,999	2.892	0.6876	12.16	0.1432
Marital Status (Ref: Married) / Not Married	3.918	1.352	11.35	0.01311
Depression (Ref: No) / Yes	0.4681	0.14	1.565	0.2114
Stress Score	1.092	1.005	1.187	0.03767

In the survey-weighted logistic regression model examining factors associated with tobacco use in the past year, marital status and stress score were significantly associated with increased odds of tobacco use. Participants who were not married had higher odds of using tobacco compared with those who were married (OR = 3.92, 95% CI: 1.35–11.35, p = 0.013). Additionally, each one-unit increase in stress score was associated with a modest but statistically significant increase in the odds of tobacco use (OR = 1.09, 95% CI: 1.01–1.19, p = 0.038). Other sociodemographic variables, including education level and income, as well as depression status, were not significantly associated with tobacco use in this model, although some point estimates suggested potential trends.

A key limitation of the present study is that some categorical variables, including marital status and income level, contained categories with small numbers of participants reporting the outcomes. To ensure model stability, these levels were combined (e.g., widowed, divorced, and separated into ‘Other/Separated’, and higher income categories collapsed), and the models report overall associations rather than estimates for each individual category. Consequently, the odds ratios presented represent average effects across all levels of each predictor, potentially obscuring meaningful differences between specific subgroups. Standard Hosmer-Lemeshow tests were not applied because they are not appropriate for complex survey designs with weights, strata, and clustering. Future studies with larger samples per category may clarify these individual effects.

5 Conclusions

In summary, this study highlights the complex interplay between sociodemographic and psychosocial factors and substance use among the target population. Marital status and stress levels emerged as significant predictors of tobacco use, suggesting that pregnant women who are not married or who experience higher stress are at greater risk. Other factors, including education, income, and depression, showed non-significant associations, indicating that their effects may be context-dependent or moderated by other variables. These findings underscore the importance of addressing psychosocial stressors in public health interventions aimed at reducing tobacco use among pregnant women and potentially other substance use behaviors. Overall, the results provide evidence for targeted prevention strategies that consider both social and psychological dimensions of risk.