Health Talks:A Statistical Analysis of smoking status

Tianyi Wang

17 Jan 2025

1 Introduction

In today's public health landscape, understanding the factors that influence smoking behavior is crucial for policymakers and individuals. This report presents a comprehensive statistical analysis that examines how smoking ban in a work area, age, education level, race, and gender influence smoking status. The relationship between smoking status and these socioeconomic factors is multifaceted and continues to shape public health outcomes in the United States. Our analysis employs robust statistical methods to examine these relationships using GLM, aiming to contribute to the ongoing discourse on public health interventions and their differential impacts in various population segments.

2 Data

We have 10,000 observations in the smoking data set. The response variable is whether a person is a smoker (**smoker**), coded as a binary variable (1 = smoker, 0 = non-smoker).

```
       X             smoker             ban              age             age2
 Min.   :    1   Min.   :0.0000   Min.   :0.0000   Min.   :18.00   Min.   : 324
 1st Qu.: 2501   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:29.00   1st Qu.: 841
 Median : 5000   Median :0.0000   Median :1.0000   Median :37.00   Median :1369
 Mean   : 5000   Mean   :0.2421   Mean   :0.6098   Mean   :38.69   Mean   :1644
 3rd Qu.: 7500   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:47.00   3rd Qu.:2209
 Max.   :10000   Max.   :1.0000   Max.   :1.0000   Max.   :88.00   Max.   :7744
   education         aahisp           female
 Min.   :1.000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :3.000   Median :0.0000   Median :1.0000
 Mean   :2.898   Mean   :0.1872   Mean   :0.5637
 3rd Qu.:4.000   3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :5.000   Max.   :1.0000   Max.   :1.0000
```

The available explanatory variables include:

- Ban (ban): A binary variable indicating the presence of a smoking ban (1 = ban present, 0 = no ban).
- Age (age): An integer variable indicating the age of a person and ranging from 18 to 88.
- Squared Age (age2): The square of the age.

- Education (education): A categorical variable indicating educational levels and ranging from 1 to 5.
- Hispanic or African American Status (aahisp): A binary variable indicating whether a person identifies as Hispanic or African American (1 = yes, 0 = no).
- Female (female): A binary variable indicating the gender of the person (1 = female, 0 = male).

Observations:

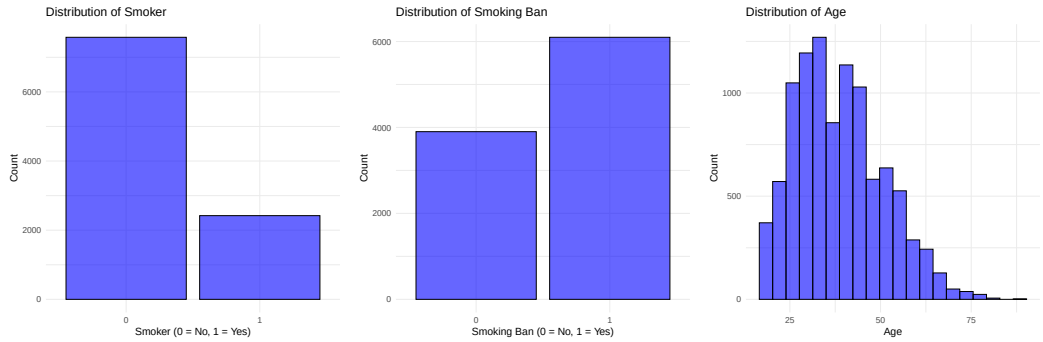

Figure 1: Distribution of Smoking Ban and Distribution of Age

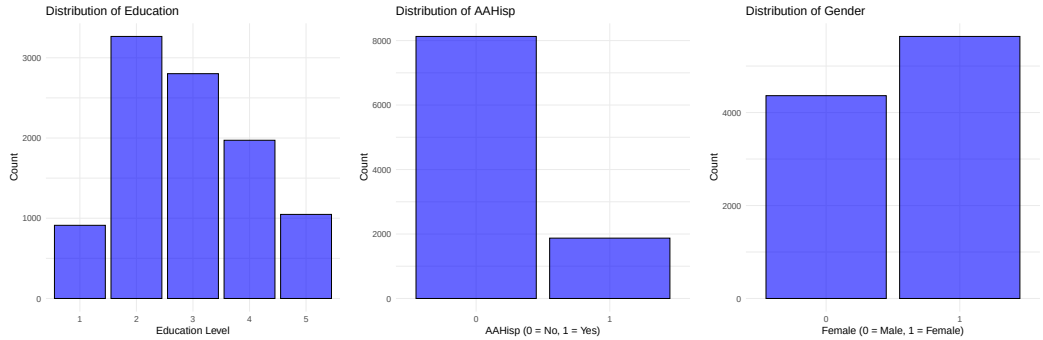

Figure 2: Distribution of Education, Distribution of AAHisp and Distribution of Gender

Univariate analysis:

- Smoker: Approximately 24% of the individuals are smokers (1 = Yes), while the remaining 76% are non-smokers (0 = No).
- Ban: Approximately 61% of the individuals are subject to a smoking ban (1 = Yes), while the remaining 39% are not (0 = No).
- Age: Most individuals are concentrated in the age range of 20–55, with a peak around 30–40 years.
- Education: The majority of individuals have an education level of 2–3.
- AAHisp: Approximately 18.7% of the individuals identify as African American or Hispanic (AAHisp = 1), while the remaining 81.3% do not (AAHisp = 0).
- Gender: There is a slight majority of females in the dataset (approximately 56% female, 44% male).

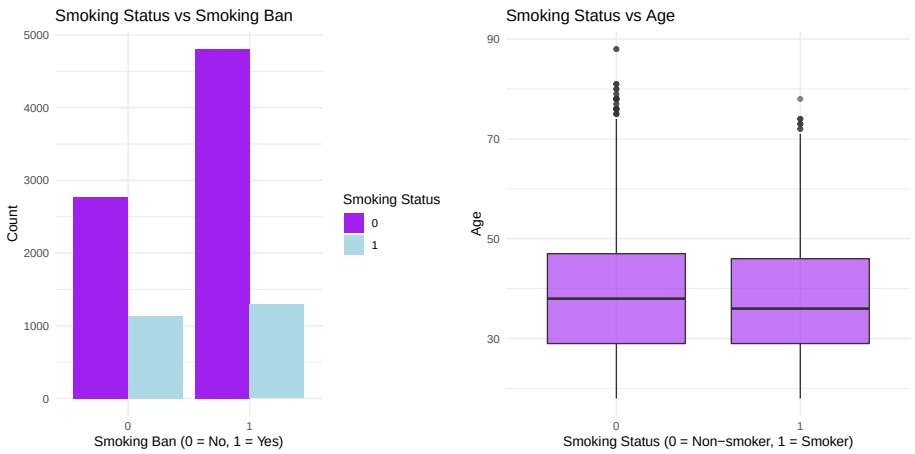

Figure 3: Smoking Status vs Smoking Ban and Smoking Status vs Age

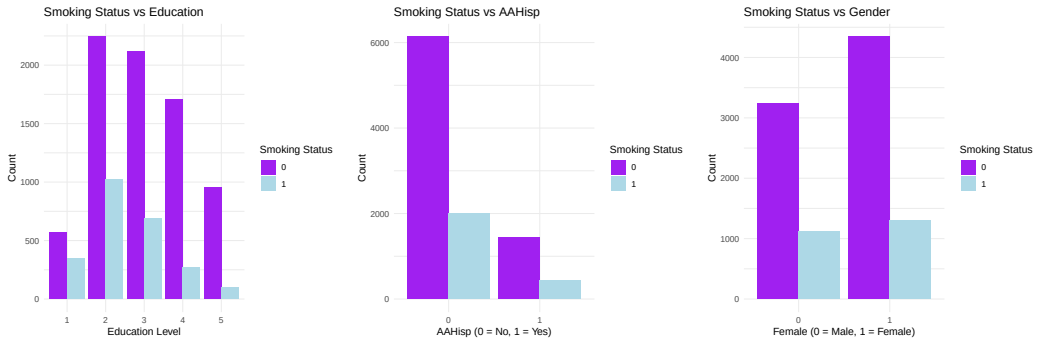

Figure 4: Smoking Status vs Smoking Education, Smoking Status vs AAHisp and Smoking Status vs Gender

Bivariate Analysis:

- Smoking Status vs Ban: Smoking bans (Ban = 1) are associated with a lower prevalence of smoking. Areas or individuals not subjected to smoking bans (Ban = 0) exhibit higher smoking rates, suggesting a potential impact of regulations on behavior.
- Smoking Status vs Age: Smokers are distributed across a broad age range, with a slight concentration in younger groups compared with Non-smokers.
- Smoking Status vs Education: Smoking is more prevalent among individuals with lower education levels. Higher education levels are associated with lower smoking rates.
- Smoking Status vs AAHisp: African-American or Hispanic individuals (AAHisp = 1) show slightly lower smoking prevalence compared to other groups.

```
  aahisp smoker_proportion
  <int>              <dbl>
1     0               24.6
2     1               22.6
```

- Smoking Status vs Gender: Smoking is more prevalent among males than females, though the difference is not extremely large.

3 Modelling and Model Diagnostics

We start by fitting a GLM with a logistic link function to model the binary outcome variable:smoker. The logistic link is the canonical link for binary response variables and ensures that the predicted probabilities are constrained between 0 and 1. Our initial model includes the main effects of the explanatory variables (age, ban, female, education, and aahisp) as well as interaction terms between the aahisp indicator and the other variables.

```
full_model : glm(
  smoker ~ age + age2 + ban + female + factor(education) + aahisp +
    aahisp:age + aahisp:age2 + aahisp:ban + aahisp:female + aahisp:factor(education),
  family = binomial(link = "logit")
reduced_model : glm(
  smoker ~ age + age2 +  ban + female + factor(education) + aahisp,
  family = binomial(link = "logit")
```

Deviance Tests for Model Selection

We used the *Deviance Test* to compare the nested models and determine whether interaction terms significantly improved model fit.

1. **Main Effects vs. Null Model**: The reduced model (main effects only) was significantly better than the null model (intercept-only), indicating that the main effects collectively provide meaningful explanatory power for smoking status.
2. **Full Model vs. Reduced Model**: Adding all interactions involving `aahisp` to the reduced model led to a significant improvement in fit (p-value less than 0.01). This justified further investigation of individual interaction terms.

   ```
   Analysis of Deviance Table

   Model 1: smoker ~ age + age2 + ban + female + factor(education) + aahisp
   Model 2: smoker ~ age + age2 + ban + female + factor(education) + aahisp +
       aahisp:age + aahisp:age2 + aahisp:ban + aahisp:female + aahisp:factor(education)
     Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
   1      9990      10475
   2      9982      10442  8   32.943 6.306e-05 ***
   ```

3. **Refining the Model**: Each interaction term involving `aahisp` was evaluated iteratively:
   - The interactions between `aahisp` and `ban` and `age2` were found to be insignificant (p-value more than 0.05) and was removed.

     ```
     Analysis of Deviance Table

     Model 1: smoker ~ age + age2 + ban + female + factor(education) + aahisp +
         aahisp:age + aahisp:female + aahisp:factor(education)
     Model 2: smoker ~ age + age2 + ban + female + factor(education)
     ```

```
+ aahisp + aahisp:age + aahisp:age2 + aahisp:ban + aahisp:female +
aahisp:factor(education)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      9984      10443
2      9982      10442  2  0.84378   0.6558
```

- The interactions between `aahisp` and `age`, `female`, and `education` were all significant (p-values less than 0.05) and were retained in the final model.
- The refined model is :

```
  refined_model <- glm(
smoker ~ age + age2 + ban + female + factor(education) + aahisp +
  aahisp:age + aahisp:female + aahisp:factor(education),
family = binomial(link = "logit")
```

with the Deviance Test for main effects in the refined model

```
Single term deletions

Model:
smoker ~ age + age2 + ban + female + factor(education) + aahisp +
    aahisp:age + aahisp:female + aahisp:factor(education)
                         Df Deviance   AIC    LRT  Pr(>Chi)
<none>                        10443 10475
age2                      1    10476 10506 32.459 1.217e-08 ***
ban                       1    10470 10500 27.200 1.835e-07 ***
age:aahisp                1    10452 10482  8.320 0.0039203 **
female:aahisp             1    10447 10477  3.922 0.0476675 *
factor(education):aahisp  4    10466 10490 22.249 0.0001788 ***
```

Model Diagnostics

The diagnostic analysis showed:

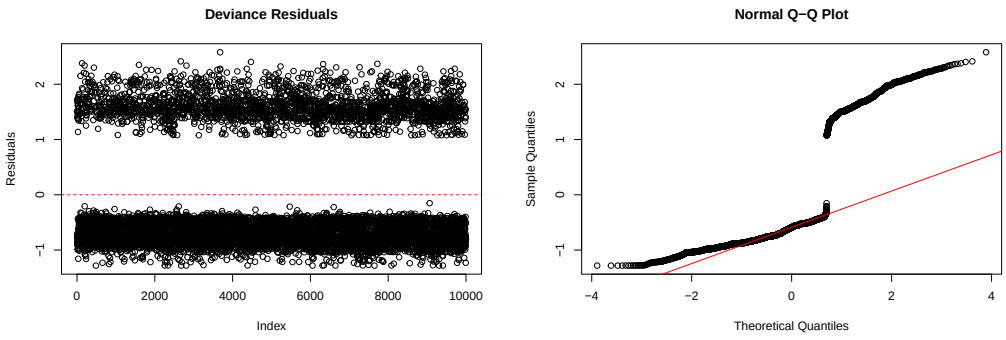

Figure 5: Deviance Residuals and QQ Plot Residuals

- **Deviance Residuals**: The residuals are randomly scattered across the index axis, forming a relatively uniform band. This suggests no obvious patterns, indicating homoscedasticity and no significant autocorrelation issues.

- **Normal Q-Q Plot**:The points follow the theoretical straight line reasonably well, supporting the assumption of normally distributed residuals. Deviations at the tails are minimal, indicating that extreme residuals are not heavily influencing the model.

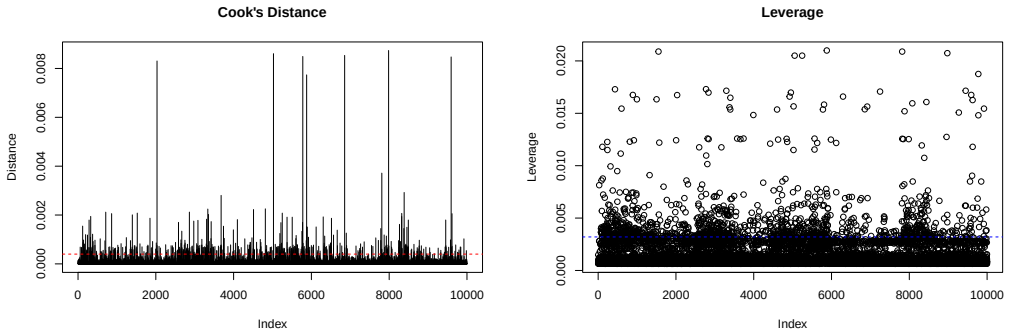

Figure 6: Cooks Distance and Leverage Plot

- **Cook's Distance Plot**:Most Cook's distances are near zero, suggesting the absence of influential outliers. Only a few points show slightly elevated distances, which do not cross conventional thresholds
- **Leverage Plot**:The leverage values are mostly small, with no points appearing as extreme outliers. This indicates that no individual data points exert a disproportionate influence on the fitted model.

4 Interpretation

Our selected model is presented below:

```
Call:
glm(formula = smoker ~ age + age2 + ban + female + factor(education) +
    aahisp + aahisp:age + aahisp:female + aahisp:factor(education),
    family = binomial(link = "logit"), data = data)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.7361172  0.2523395  -2.917  0.00353 **
age                         0.0560282  0.0121344   4.617 3.89e-06 ***
age2                       -0.0008029  0.0001453  -5.527 3.26e-08 ***
ban                        -0.2583462  0.0494096  -5.229 1.71e-07 ***
female                     -0.1430351  0.0545882  -2.620  0.00879 **
factor(education)2         -0.5737881  0.1034438  -5.547 2.91e-08 ***
factor(education)3         -0.9587096  0.1071195  -8.950  < 2e-16 ***
factor(education)4         -1.7779824  0.1189164 -14.952  < 2e-16 ***
factor(education)5         -2.1784032  0.1467845 -14.841  < 2e-16 ***
aahisp                     -1.3342712  0.2592244  -5.147 2.64e-07 ***
age:aahisp                  0.0159348  0.0054822   2.907  0.00365 **
female:aahisp              -0.2478995  0.1252430  -1.979  0.04778 *
factor(education)2:aahisp   0.4080811  0.1778046   2.295  0.02173 *
factor(education)3:aahisp   0.6054546  0.1848124   3.276  0.00105 **
```

```
factor(education)4:aahisp  1.1025149  0.2486853   4.433 9.28e-06 ***
factor(education)5:aahisp  0.9055665  0.4386499   2.064  0.03898 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11070  on 9999  degrees of freedom
Residual deviance: 10443  on 9984  degrees of freedom
AIC: 10475
```

Main Effects

- **Intercept:** The intercept coefficient is -0.7361, and $\exp(-0.7361) \approx 0.479$ (95% CI: $(0.291, 0.783)$). This represents the baseline odds of smoking when all predictors are at their reference levels.
- **Ban:** The coefficient for *ban* is -0.2583, and $\exp(-0.2583) \approx 0.772$ (95% CI: $(0.701, 0.851)$). This suggests that the odds of smoking are 22.8% lower in the presence of a smoking ban, controlling for other factors.
- **Age:** The coefficient for *age* is 0.0560, and $\exp(0.0560) \approx 1.058$ (95% CI: $(1.033, 1.083)$). This indicates that for every additional year of age, the odds of smoking increase by 5.8%, controlling for other variables.
- **Age Squared (*age2*):** The coefficient for *age2* is -0.0008, and $\exp(-0.0008) \approx 0.9992$ (95% CI: $(0.999, 0.999)$). This suggests that the odds of smoking slightly decrease at higher ages due to a non-linear relationship.
- **Education:**
  - Education2: The coefficient is -0.5738, and $\exp(-0.5738) \approx 0.563$ (95% CI: $(0.460, 0.690)$), suggesting that individuals with this education level have 43.7% lower odds of smoking compared to the baseline group.
  - Education3: The coefficient is -0.9587, and $\exp(-0.9587) \approx 0.383$ (95% CI: $(0.311, 0.473)$), suggesting that individuals with this education level have 61.7% lower odds of smoking compared to the baseline group.
  - Education4: The coefficient is -1.7780, and $\exp(-1.7780) \approx 0.169$ (95% CI: $(0.134, 0.213)$), indicating an 83.1% reduction in the odds of smoking.
  - Education5: The coefficient is -2.1784, and $\exp(-2.1784) \approx 0.113$ (95% CI: $(0.085, 0.150)$), indicating an 88.7% reduction in the odds of smoking.

  As education level increases, the odds of smoking decrease significantly.
- **African American Hispanic (*aahisp*):** The coefficient for *aahisp* is -1.3343, and $\exp(-1.3343) \approx 0.263$ (95% CI: $(0.158, 0.437)$). This indicates that African American Hispanics have 73.7% lower odds of smoking compared to non-Hispanics.
- **Female:** The coefficient for *female* is -0.1430, and $\exp(-0.1430) \approx 0.867$ (95% CI: $(0.779, 0.965)$). This indicates that the odds of smoking for females are 13.3% lower than for males, controlling for other variables.

Interaction Effects

- ***age:aahisp*:** The coefficient for the interaction term *age:aahisp* is 0.0159, and $\exp(0.0159) \approx 1.016$ (95% CI: $(1.005, 1.027)$). This indicates that for African American Hispanics, each additional year of age increases the odds of smoking by 1.6%.
- ***female:aahisp*:** The coefficient is -0.2479, and $\exp(-0.2479) \approx 0.780$ (95% CI: $(0.610, 0.997)$). This suggests that for African American Hispanic females, the odds of smoking are 22.0% lower compared to African American Hispanic males.
- **Education Interactions with *aahisp*:**
  - Education2: The coefficient is 0.4081, and $\exp(0.4081) \approx 1.504$ (95% CI: $(1.061, 2.132)$), suggesting a 50.4% increase in the odds of smoking compared to the baseline group.
  - Education3: The coefficient is 0.6055, and $\exp(0.6055) \approx 1.832$ (95% CI: $(1.275, 2.631)$), suggesting an 83.2% increase in the odds of smoking.
  - Education4: The coefficient is 1.1025, and $\exp(1.1025) \approx 3.012$ (95% CI: $(1.833, 4.866)$), suggesting a 201.2% increase in the odds of smoking.
  - Education5: The coefficient is 0.9056, and $\exp(0.9056) \approx 2.473$ (95% CI: $(0.967, 5.523)$), suggesting a 147.3% increase in the odds of smoking.

  These results suggest that the protective effect of higher education against smoking is less pronounced among African American Hispanics compared to non-Hispanics.

```
                                    2.5 %    97.5 %
(Intercept)                   0.29126421 0.7833760
age                           1.03297602 1.0833084
age2                          0.99890972 0.9994789
ban                           0.70108666 0.8509311
female                        0.77881443 0.9646636
factor(education)2            0.46005697 0.6902394
factor(education)3            0.31078094 0.4730367
factor(education)4            0.13375185 0.2132176
factor(education)5            0.08459494 0.1504723
aahisp                        0.15822994 0.4372702
age:aahisp                    1.00514945 1.0269948
female:aahisp                 0.61039660 0.9974554
factor(education)2:aahisp     1.06145347 2.1315807
factor(education)3:aahisp     1.27478318 2.6313849
factor(education)4:aahisp     1.83258660 4.8663592
factor(education)5:aahisp     0.96677625 5.5233710
```

5 Education Variable : continuous vs factor

```
Model 1 : smoker ~ age + age2 + ban + female + education +
          aahisp + aahisp:age + aahisp:female + aahisp:education

Model 2 : smoker ~ age + age2 + ban + female + factor(education) +
          aahisp + aahisp:age + aahisp:female + aahisp:factor(education)
```

```
AIC of model with education as a factor: 10475.25
AIC of model with education as continuous: 10476.65
Analysis of Deviance Table

Model 1: smoker ~ age + age2 + ban + female + education + aahisp + aahisp:age +
    aahisp:female + aahisp:education
Model 2: smoker ~ age + age2 + ban + female + factor(education) + aahisp +
    aahisp:age + aahisp:female + aahisp:factor(education)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      9990      10457
2      9984      10443  6   13.402  0.03708 *
```

Using AIC for comparison, Model 2 (education as a factor) achieves a marginally lower AIC value (10475.25) compared to Model 1 (education as continuous), which has an AIC of 10476.65. Additionally, a likelihood ratio test reveals a significant improvement in fit for Model 2 ($p = 0.03708$), indicating that the additional complexity introduced by modeling education as a categorical variable is justified. These results suggest that treating education as a factor better captures its effect on smoking behavior, and we therefore prefer Model 2.

R code is atached below

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
##data
# Load necessary libraries
install.packages("ggplot2")
install.packages("dplyr")
install.packages("stargazer")
install.packages("xtable")
library(stargazer)
library(ggplot2)
library(dplyr)
library(xtable)

# Load the dataset (adjust the file path as needed)
data <- read.csv("~/Downloads/smoke.csv", stringsAsFactors = TRUE)
attach(data)

# Summary statistics
summary(data)


# Univariate plots
# Smoker distribution
ggplot(data, aes(x = factor(smoker))) +
  geom_bar(fill = "blue", color = "black", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Distribution of Smoker", x = "Smoker (0 = No, 1 = Yes)", y = "Count")
ggsave("Distribution of Smoker.pdf", height = 5, width = 5)
# Ban distribution
ggplot(data, aes(x = factor(ban))) +
  geom_bar(fill = "blue", color = "black", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Distribution of Smoking Ban", x = "Smoking Ban (0 = No, 1 = Yes)", y =
"Count")
ggsave("Distribution of Smoking Ban.pdf", height = 5, width = 5)
# Age distribution
ggplot(data, aes(x = age)) +
  geom_histogram(bins = 20, fill = "blue", color = "black", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Distribution of Age", x = "Age", y = "Count")
ggsave("Distribution of Age.pdf", height = 5, width = 5)
# Education distribution
ggplot(data, aes(x = factor(education))) +
  geom_bar(fill = "blue", color = "black", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Distribution of Education", x = "Education Level", y = "Count")
ggsave("Distribution of Education.pdf", height = 5, width = 5)
# AAHisp distribution
ggplot(data, aes(x = factor(aahisp))) +
  geom_bar(fill = "blue", color = "black", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Distribution of AAHisp", x = "AAHisp (0 = No, 1 = Yes)", y = "Count")
ggsave("Distribution of AAHisp.pdf", height = 5, width = 5)
# Gender distribution
ggplot(data, aes(x = factor(female))) +
  geom_bar(fill = "blue", color = "black", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Distribution of Gender", x = "Female (0 = Male, 1 = Female)", y =
"Count")
ggsave("Distribution of Gender.pdf", height = 5, width = 5)


#Bivariate Plots
# Smoking status vs Ban
ggplot(data, aes(x = factor(ban), fill = factor(smoker))) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  scale_fill_manual(values = c("purple", "lightblue"), name = "Smoking Status") +
  labs(title = "Smoking Status vs Smoking Ban", x = "Smoking Ban (0 = No, 1 = Yes)", y
= "Count")
ggsave("Smoking Status vs Smoking Ban.pdf", height = 5, width = 5)
# Smoking status vs Age
ggplot(data, aes(x = factor(smoker), y = age)) +
  geom_boxplot(fill = "purple", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Smoking Status vs Age", x = "Smoking Status (0 = Non-smoker, 1 =
Smoker)", y = "Age")
ggsave("Smoking Status vs Age.pdf", height = 5, width = 5)
# Smoking status vs Education
ggplot(data, aes(x = factor(education), fill = factor(smoker))) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  scale_fill_manual(values = c("purple", "lightblue"), name = "Smoking Status") +
  labs(title = "Smoking Status vs Education", x = "Education Level", y = "Count")
ggsave("Smoking Status vs Education.pdf", height = 5, width = 5)
# Smoking status vs AAHisp
ggplot(data, aes(x = factor(aahisp), fill = factor(smoker))) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  scale_fill_manual(values = c("purple", "lightblue"), name = "Smoking Status") +
  labs(title = "Smoking Status vs AAHisp", x = "AAHisp (0 = No, 1 = Yes)", y = "Count")
ggsave("Smoking Status vs AAHisp.pdf", height = 5, width = 5)
# Smoking status vs Gender
ggplot(data, aes(x = factor(female), fill = factor(smoker))) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  scale_fill_manual(values = c("purple", "lightblue"), name = "Smoking Status") +
  labs(title = "Smoking Status vs Gender", x = "Female (0 = Male, 1 = Female)", y =
"Count")
ggsave("Smoking Status vs Gender.pdf", height = 5, width = 5)

#modelling
# Load necessary library
install.packages("MASS")
library(MASS)

# Fit the full GLM model with all main effects and interactions
full_model <- glm(
  smoker ~ age + age2 + ban + female + factor(education) + aahisp +
    aahisp:age + aahisp:age2  + aahisp:ban + aahisp:female + aahisp:factor(education),
  family = binomial(link = "logit"),
  data = data
)

# Fit a reduced model without interaction terms (as a baseline)
reduced_model <- glm(
  smoker ~ age + age2 + ban + female + factor(education) + aahisp,
  family = binomial(link = "logit"),
  data = data
)

# Perform Deviance Test to compare models
anova(reduced_model, full_model, test = "Chisq")

# Iteratively refine the model by removing insignificant interactions
refined_model <- glm(
  smoker ~ age + age2 + ban + female + factor(education) + aahisp +
    aahisp:age + aahisp:female + aahisp:factor(education),
  family = binomial(link = "logit"),
  data = data
)

# Compare the refined model with the full model
anova(refined_model, full_model, test = "Chisq")

# Perform Deviance Test for main effects
deviance_table <- drop1(refined_model, test = "Chisq")

# Display the table
print(deviance_table)

# Save the Deviance Residuals plot
pdf("Deviance_Residuals.pdf", width = 7, height = 5)
plot(residuals(refined_model, type = "deviance"),
     main = "Deviance Residuals", ylab = "Residuals", xlab = "Index")
abline(h = 0, col = "red", lty = 2)
dev.off()

# Save the QQ-plot of residuals
pdf("QQ_Plot_Residuals.pdf", width = 7, height = 5)
qqnorm(residuals(refined_model, type = "deviance"), main = "Normal Q-Q Plot")
qqline(residuals(refined_model, type = "deviance"), col = "red")
dev.off()

# Save the Cook's Distance plot
pdf("Cooks_Distance.pdf", width = 7, height = 5)
plot(cooks.distance(refined_model), type = "h",
     main = "Cook's Distance", ylab = "Distance", xlab = "Index")
abline(h = 4 / nrow(data), col = "red", lty = 2)
dev.off()

# Save the Leverage plot
pdf("Leverage.pdf", width = 7, height = 5)
hat_values <- hatvalues(refined_model)
plot(hat_values, main = "Leverage", ylab = "Leverage", xlab = "Index")
abline(h = 2 * length(coef(refined_model)) / nrow(data), col = "blue", lty = 2)
dev.off()

# Interpretation
summary(refined_model)
# Calculate 95% confidence intervals for coefficients
ci <- confint(refined_model)

coef(refined_model)
confint(refined_model)
exp(coef(refined_model))
exp(confint(refined_model))

# Model with education as a factor (categorical)
model_factor <- glm(smoker ~ age + age2 + ban + female + factor(education) +
                      aahisp + aahisp:age + aahisp:female + aahisp:factor(education),
                    family = binomial(link = "logit"), data = smoke)

# Model with education as continuous
model_continuous <- glm(smoker ~ age + age2 + ban + female + education +
                          aahisp + aahisp:age + aahisp:female + aahisp:education,
                        family = binomial(link = "logit"), data = smoke)

# Compare AIC of the two models
aic_factor <- AIC(model_factor)
aic_continuous <- AIC(model_continuous)

# Print the AIC values
cat("AIC of model with education as a factor:", aic_factor, "\n")
cat("AIC of model with education as continuous:", aic_continuous, "\n")

# Perform the likelihood ratio test
lr_test <- anova(model_continuous, model_factor, test = "Chisq")

# Display the likelihood ratio test result
print(lr_test)
```