Exploration of the Influential Factor of the Presidential Election Results in 2008 Using Logistic Regression Models

Mike Liu, Osvaldo Hernandez-Segura, and Daisy Yu

2024-12-05

```
library(Stat2Data)
library(ggplot2)
library(dplyr)
library(car)
data("Election08")
```

Abstract

This study investigates state-level factors to predict whether Barack Obama or John McCain won the 2008 U.S. presidential election. Using both simple and multiple logistic regression models, we analyze predictors such as income, education levels, and political leaning to determine their relationship with election outcomes. Empirical logit plots are employed to verify the linearity assumption for logistic regression, ensuring the validity of the models. Our analysis identifies the most significant contributors to predicting election results and evaluates the effectiveness of the multiple logistic regression model compared to simpler models. The findings provide insights into the factors most influential in determining the outcome of the 2008 U.S. presidential election.

Introduction

This study investigates the state-level socioeconomic and political factors that most effectively predicted the outcomes of the 2008 U.S. presidential election using the Election08 dataset. By applying logistic regression models, we aim to explore the relationships between predictors such as per capita income, educational attainment, and political leaning and their impact on election results. This analysis seeks to identify the most influential factors driving voting behavior across the 50 states and the District of Columbia.

Research Question

Which state-level factor—political leaning, education level, or income—is the most influential in predicting the outcome of the 2008 U.S. presidential election? Additionally, does a multiple logistic regression model provide better predictive performance compared to individual simple logistic regression models for this election?

Expected Findings

We anticipate that political leaning, as measured by the difference between the percentage of Democrats and Republicans in each state (Dem.Rep), will be the most influential factor in determining the outcome of the 2008 U.S. presidential election. States with a higher proportion of Democrats relative to Republicans are expected to have a higher likelihood of being won by Barack Obama. Additionally, we hypothesize that a multiple logistic regression model incorporating multiple predictors, including Income, HS, BA, and Dem.Rep, will provide the most accurate predictions of election outcomes. This is because the interaction of socioeconomic factors with political alignment likely captures more nuanced variations in voting behavior than single-variable models. The multivariable approach is expected to demonstrate the best fit and predictive power, as assessed by appropriate statistical metrics.

Context and Relevance

Understanding the demographic and socioeconomic factors that influence voting behavior is critical for political strategy and policy analysis. This study examines how variables such as income and education correlate with election outcomes, shedding light on patterns in state-level voting behavior. By focusing on the differences between Democratic and Republican political affiliations, this analysis highlights key dynamics in electoral decision-making.

Research Population

The research population consists of all 50 U.S. states and the District of Columbia, as represented in the Election08 dataset. This includes data on state-level socioeconomic characteristics (per capita income, high school, and college education levels) and political leaning (difference in Democrat and Republican support). The population is analyzed to predict whether Barack Obama or John McCain won each state in the 2008 U.S. presidential election.

Methodology

This study provides a structured approach to understanding how state-level characteristics influenced the 2008 presidential election and offers insights into broader patterns of electoral behavior in the United States. The dataset of reference contains information from all 50 states and the District of Columbia for the 2008 U.S. presidential election. This analysis consists of 5 models, 4 of which investigate if Income, HS, BA, and Dem.Rep is associated with the odds that Obama (Democrat) wins the state in 2008 (ObamaWin = 1), and 1 of which is an interaction model that investigates the joint effect of these four variables.

Rationale

The primary goal of this project is to predict state-level election outcomes for the 2008 U.S. presidential election. The dependent variable, ObamaWin, is categorical (1 = Obama won, 0 = McCain won), making logistic regression the most appropriate modeling technique. Logistic regression is specifically designed to model binary outcomes by estimating the probability of success (in this case, Obama winning a state) as a function of predictor variables.

The following modeling process is used to achieve this objective:

1. Simple Logistic Regression:

- To explore the relationship between each predictor (Income, HS, BA, Dem.Rep) and the dependent variable independently.
- This step allows us to identify which variables are individually significant predictors of the election outcome.

2. Multiple Logistic Regression:
   - To account for the interaction and combined effect of all predictors on the election outcome.
   - This model provides a more comprehensive understanding of how socioeconomic and political factors collectively influence voting patterns.
3. Model Comparison:
   - We evaluate the models using tools such as confidence intervals, G-tests and AIC values. This helps determine which model best explains the variability in the election outcomes.

Why Choose Logistic Regression?

1. Suitability for Binary Outcomes: Logistic regression is tailored explicitly for binary response variables like ObamaWin.
2. Probabilistic Interpretation: The model outputs probabilities, allowing us to estimate the likelihood of Obama winning in each state.
3. Flexibility with Predictors: Logistic regression can handle a mix of numerical predictors (for example, Income, Dem.Rep) and their potential interactions.
4. Diagnostic Tools: Logistic regression provides tools to assess model fit, assumptions, and variable importance, ensuring robust and interpretable results.

This methodological approach is comprehensive yet efficient, ensuring that our findings are both statistically valid and practically relevant for understanding the factors influencing the 2008 election outcomes.

Dataset Overview

1. Description of the Dataset: The Election08 dataset provides state-by-state information related to the 2008 U.S. presidential election. It includes seven variables, capturing key socioeconomic, educational, and political characteristics of each state alongside the election outcome. These variables enable an in-depth analysis of the factors contributing to Barack Obama's or John McCain's victories in individual states.
2. Data Variable Dictionary:

- State: Name of the state.
- Abr: Abbreviation for the state.
- Income: Per capita income in the state as of 2007 (in US dollars).
- HS: Percentage of adults with at least a high school education.
- BA: Percentage of adults with at least a college education.
- Dem.Rep: Difference in percent Democrat and percent Republican (based on a 2008 Gallup survey).
- ObamaWin:
  - 1: Obama (Democrat) won the state in 2008.
  - 0: McCain (Republican) won the state in 2008.

Dataframe Details

1. Observations: 51 (50 states plus the District of Columbia).
2. Variables: 7 (4 independent variables and 1 dependent variable).
    - Independent Variables:
        - Income: Numerical continuous variable
        - HS: Numerical continuous variable
        - BA: Numerical continuous variable
        - Dem.Rep: Numerical continuous variable
    - Dependent Variable:
        - ObamaWin: Categorical binary variable

Data Sources:

1. Income Data: U.S. Census Bureau, Table 659. Personal Income Per Capita (2007).
2. High School Education Data: U.S. Census Bureau, 1990 Census of Population, [NCES Digest (http://nces.ed.gov/programs/digest/d08/tables/dt08_011.asp).
3. College Education Data: U.S. Census Bureau, Table 225. Educational Attainment by State (2007).
4. Political Leaning Data: Gallup, [State of the States: Political Party Affiliation] (http://www.gallup.com/poll/114016/state-states-political-party-affiliation.aspx#1).

Data Collection Procedure

The dataset was compiled by merging information from the sources listed above, ensuring consistency in definitions and measurements across variables. Each variable was carefully checked to match its description in the dataset's metadata, and all data represent the state-level characteristics leading up to the 2008 election. This comprehensive and systematic approach ensures the dataset is suitable for logistic regression analysis, allowing for robust predictions of the 2008 U.S. presidential election results.

Data Summary

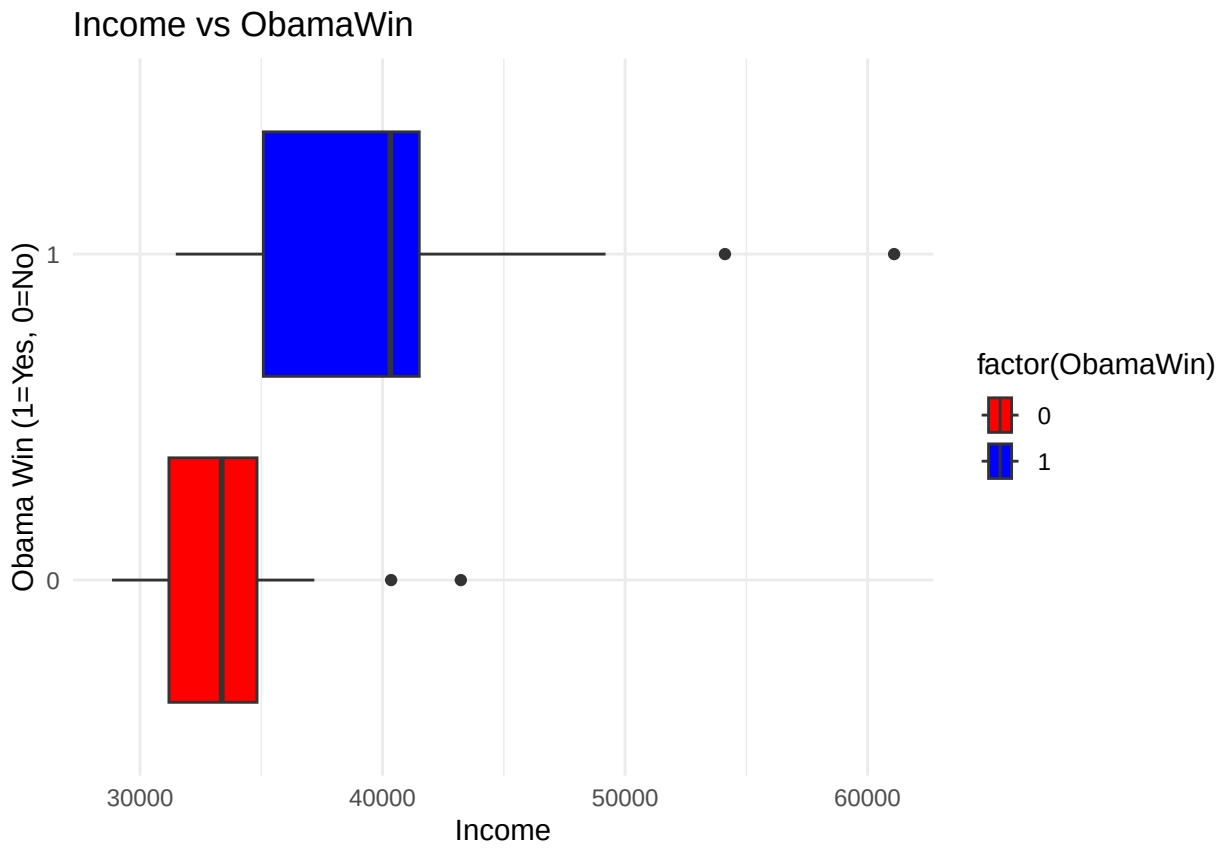
```
summary(Election08)
```

```
##          State         Abr         Income            HS              BA       
##  Alabama    : 1   AK     : 1   Min.   :28845   Min.   :78.50   Min.   :17.30  
##  Alaska     : 1   AL     : 1   1st Qu.:33536   1st Qu.:83.00   1st Qu.:24.20  
##  Arizona    : 1   AR     : 1   Median :36047   Median :87.00   Median :25.80  
##  Arkansas   : 1   AZ     : 1   Mean   :37642   Mean   :86.00   Mean   :27.15  
##  California : 1   CA     : 1   3rd Qu.:40544   3rd Qu.:89.05   3rd Qu.:29.65  
##  Colorado   : 1   CO     : 1   Max.   :61092   Max.   :91.20   Max.   :47.50  
##  (Other)    :45   (Other):45                                                  
##     Dem.Rep          ObamaWin      
##  Min.   :-23.00   Min.   :0.0000  
##  1st Qu.:  3.00   1st Qu.:0.0000  
##  Median : 12.00   Median :1.0000  
##  Mean   : 12.31   Mean   :0.5686  
```

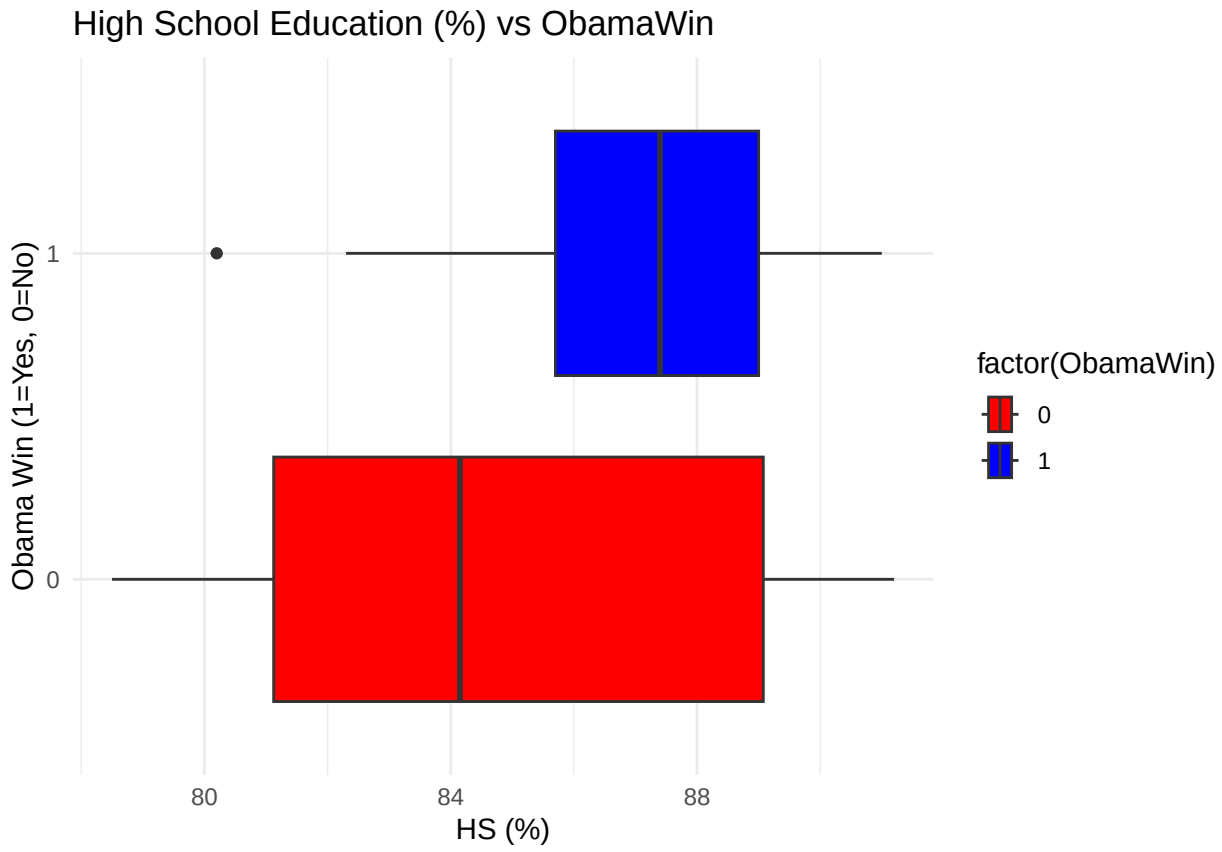
```
##  3rd Qu.: 19.00   3rd Qu.:1.0000
##  Max.   : 75.00   Max.   :1.0000
##
```

Visualization of Relationships Between Variables

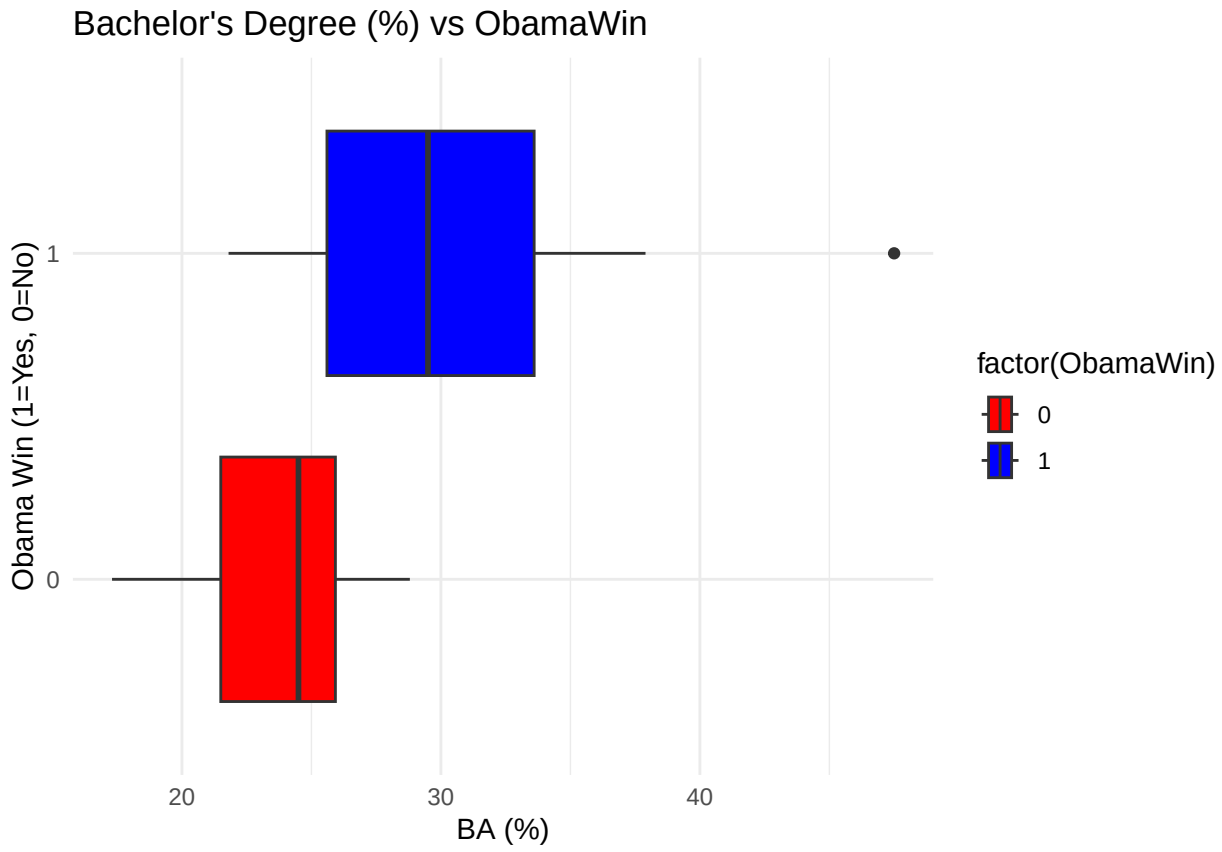
```
create_boxplot <- function(data, predictor, response, title, x_label) {
  ggplot(data, aes(x = !!sym(predictor), y = factor(!!sym(response)), fill = factor(!!sym(response)))) +
    geom_boxplot() +
    labs(title = title, x = x_label, y = "Obama Win (1=Yes, 0=No)") +
    scale_fill_manual(values = c("red", "blue")) +
    theme_minimal()
}
create_boxplot(Election08, "Income", "ObamaWin", "Income vs ObamaWin", "Income")
```



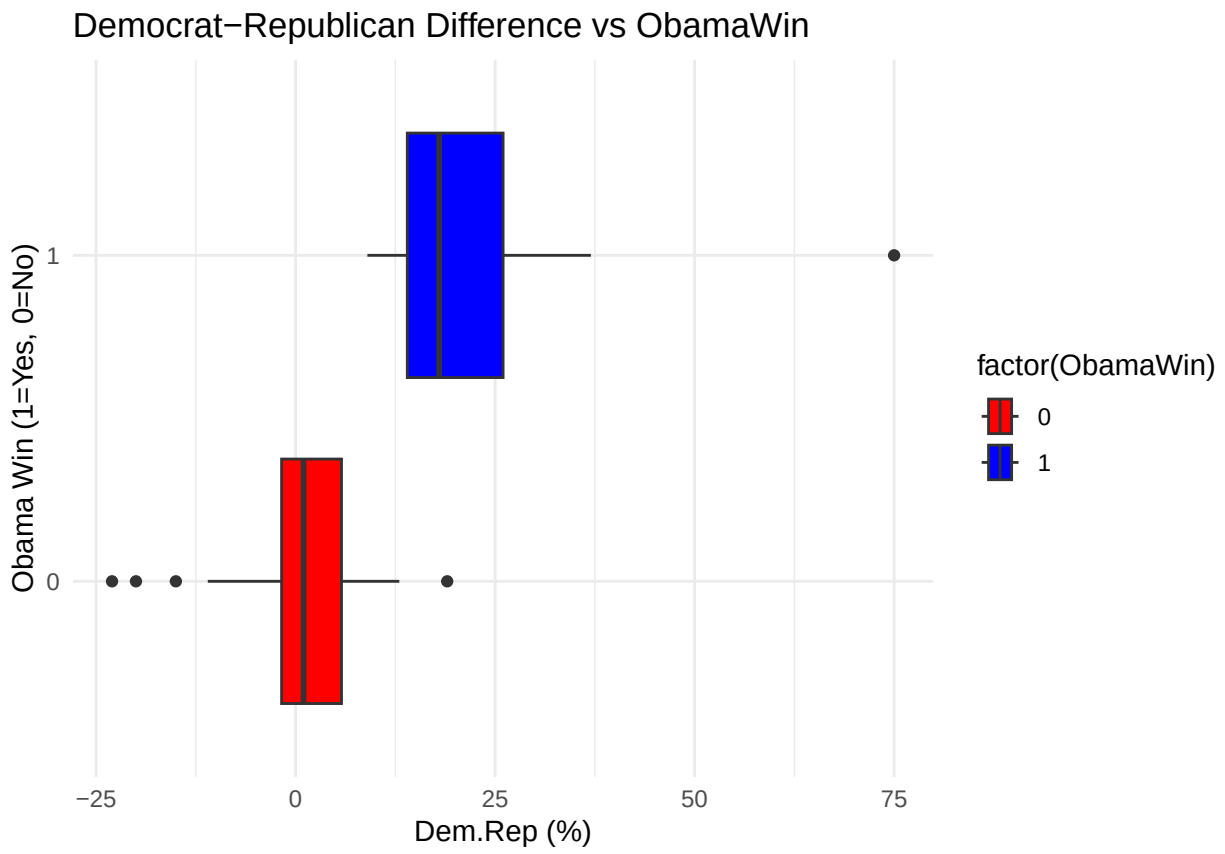

```
create_boxplot(Election08, "HS", "ObamaWin", "High School Education (%) vs ObamaWin", "HS (%)")
```




```
create_boxplot(Election08, "BA", "ObamaWin", "Bachelor's Degree (%) vs ObamaWin", "BA (%)")
```




```
create_boxplot(Election08, "Dem.Rep", "ObamaWin", "Democrat-Republican Difference vs ObamaWin", "Dem.Rep
```



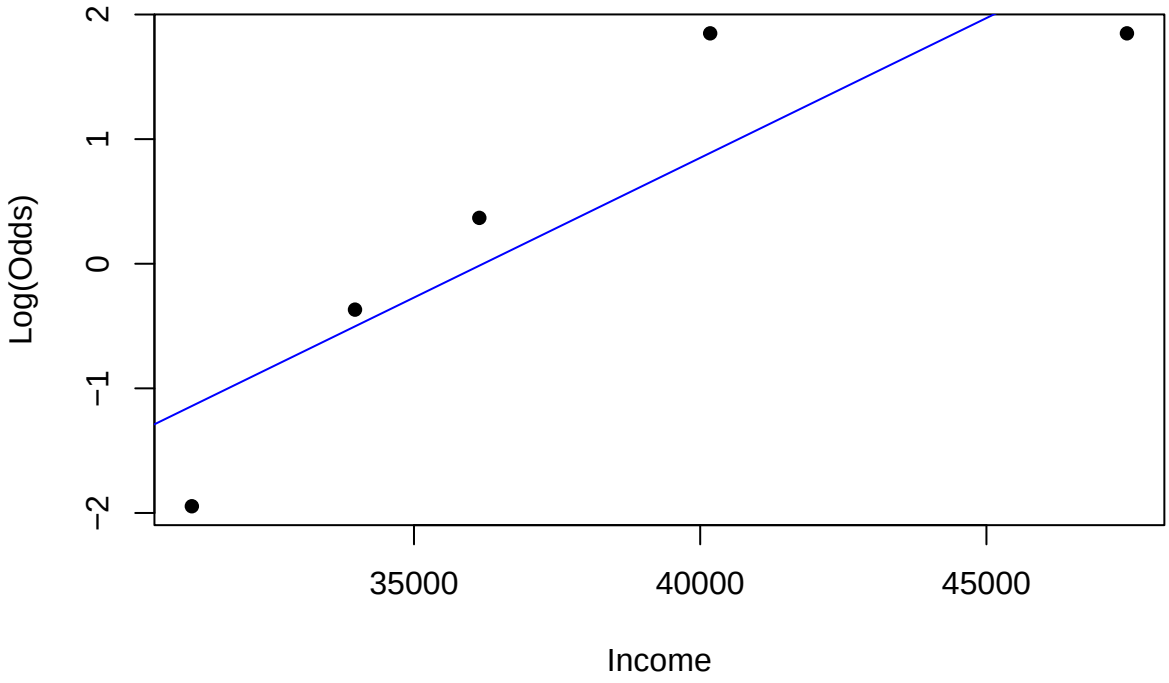
Based on the first three boxplots regarding average income and educatuion levels, we can see that among the states that Obama won, people in these states tend to have a higher average income and education level. Based on the boxplot of democratic representation, there are also more people supporting the democratic party in the states that Obama won.

Statistical Analysis

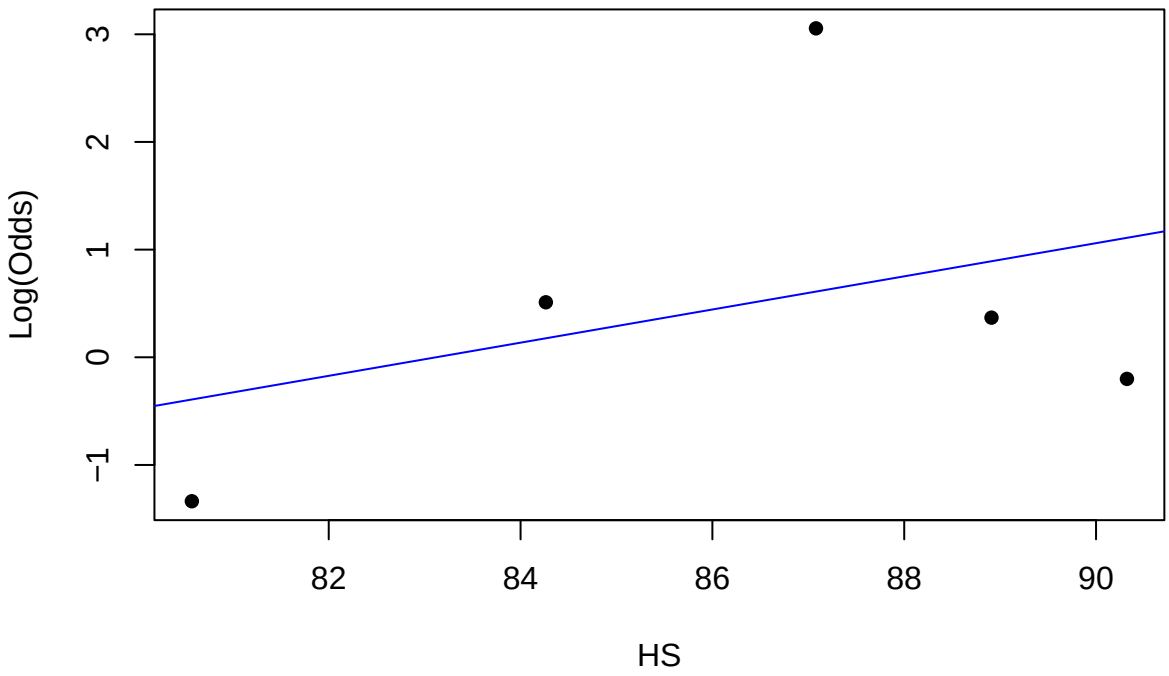
Check Conditions

The conditions for logistic regressions are linearity, randomness, and independence. The linearity condition means that a linear relationship between the log(odds) and the predictor should exist. Also, we assume that the process of data collection is random and that each observation is independent, following a Bernoulli process.

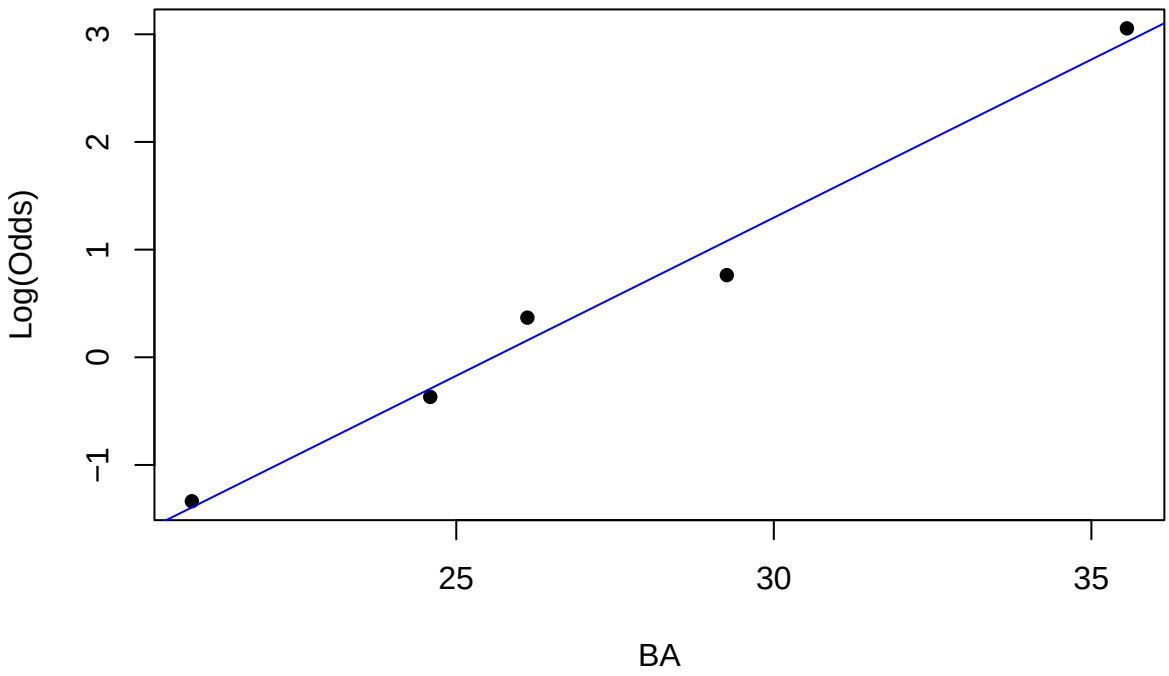
```
emplogitplot1(ObamaWin ~ Income, data = Election08, ngroups = 5)
```



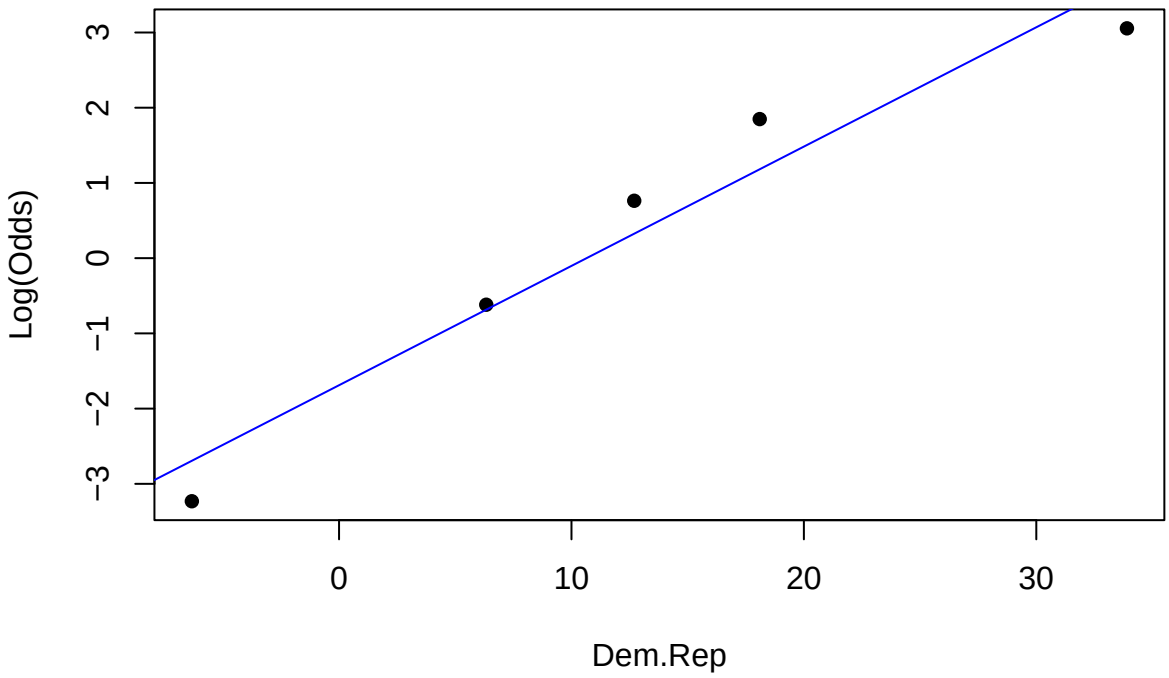

```
emplogitplot1(ObamaWin ~ HS, data = Election08, ngroups = 5)
```




```
emplogitplot1(ObamaWin ~ BA, data = Election08, ngroups = 5)
```




```
emplogitplot1(ObamaWin ~ Dem.Rep, data = Election08, ngroups = 5)
```



Thus, the linearity condition was assessed using empirical logit plots, which confirmed an approximately linear relationship between the predictors (Income, HS, BA, Dem.Rep) and the log odds of the election outcome (ObamaWin). This supports the validity of the logistic regression model.

Correlation Matrix & Variance Inflation Factor (VIF)

```
cor_matrix <- cor(Election08[, c("Income", "HS", "BA", "Dem.Rep")])
print(cor_matrix)
```

```
##             Income          HS        BA     Dem.Rep
## Income  1.0000000  0.29074115 0.8276107  0.59587400
## HS      0.2907412  1.00000000 0.4383784 -0.03188233
## BA      0.8276107  0.43837840 1.0000000  0.58496075
## Dem.Rep 0.5958740 -0.03188233 0.5849607  1.00000000
```

```
vif(lm(Dem.Rep ~ Income + HS + BA, data = Election08))
```

```
##   Income       HS       BA
## 3.240109 1.263679 3.671864
```

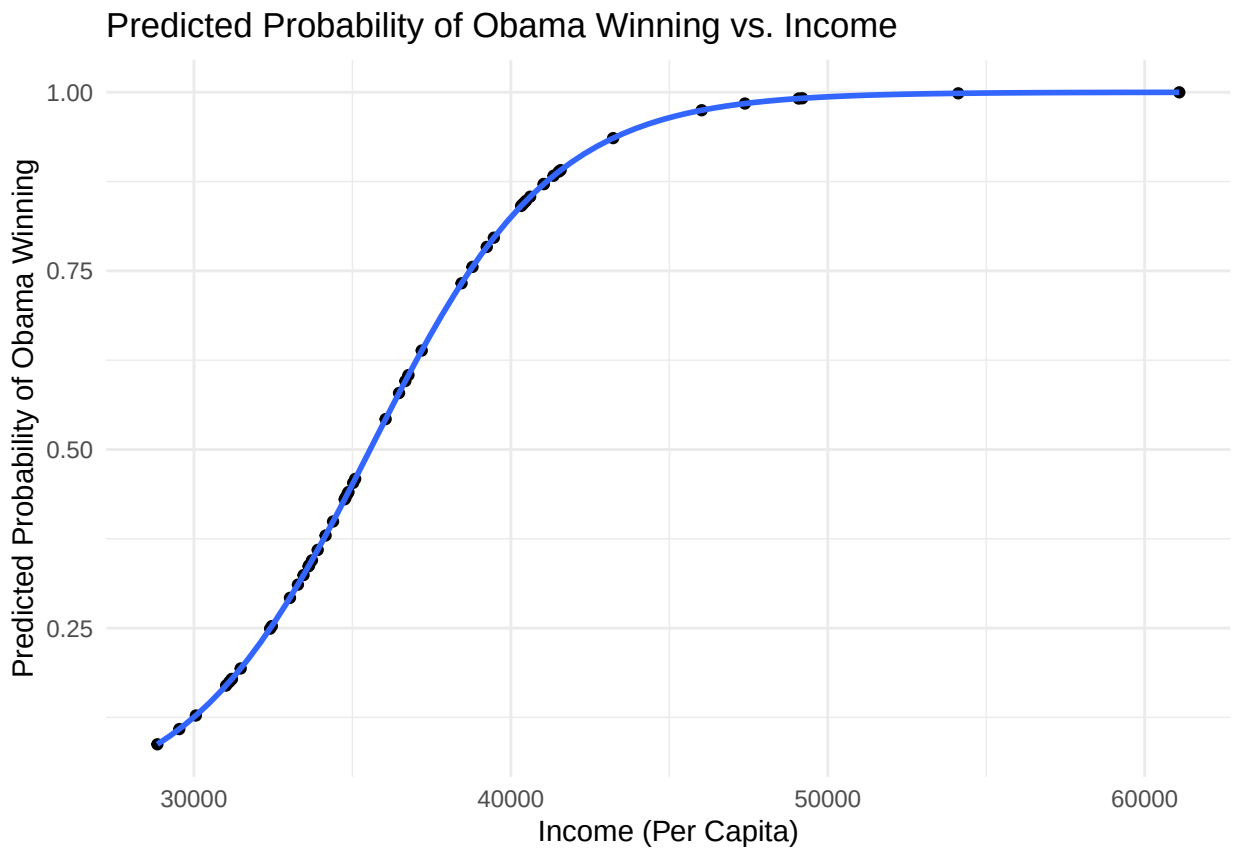
The correlation matrix revealed a strong positive correlation between Income and BA (r=0.8276), indicating a potential overlap in these predictors. However, the VIF values for all predictors were below 5, confirming that multicollinearity is not a significant concern. Still, it is worth noting that Income and BA may share similar explanatory power in the model.

Simple Logistic Regression Model Construction

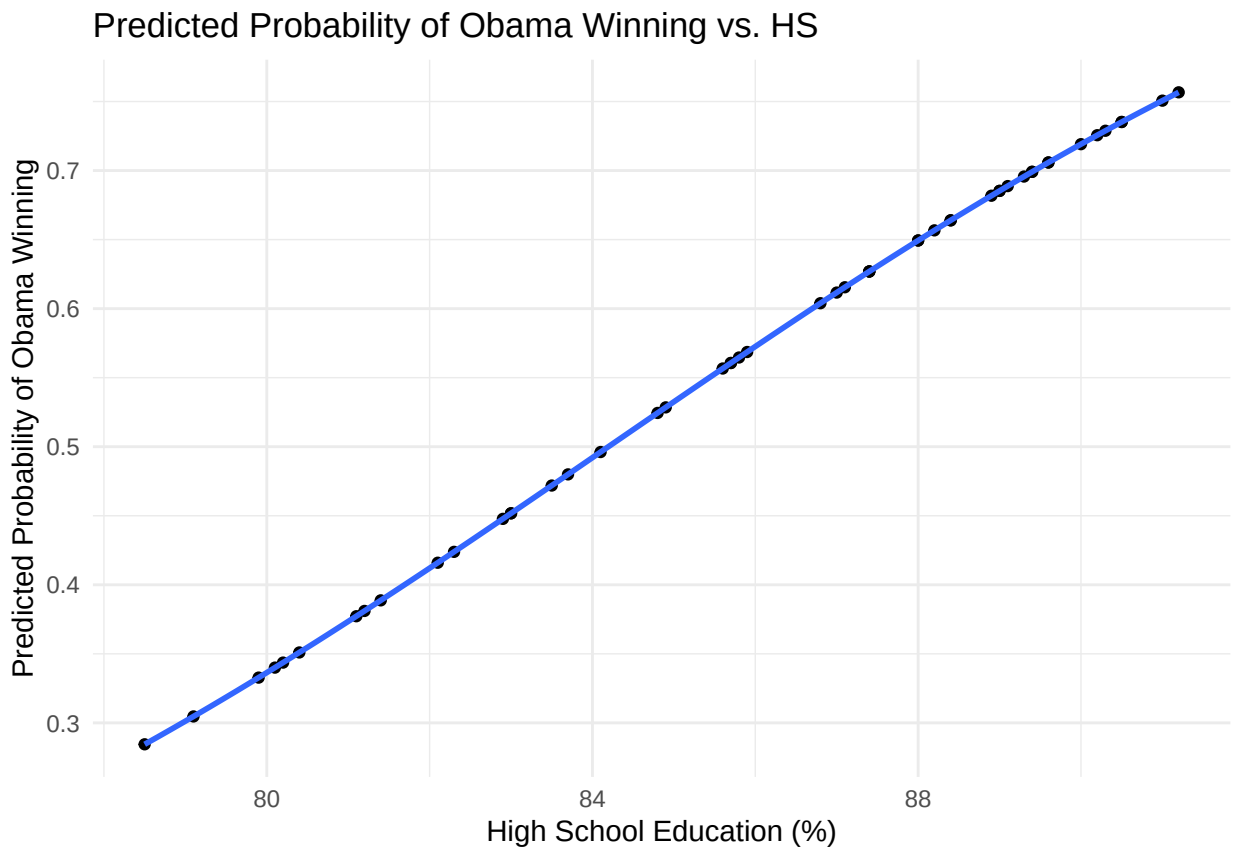
```
model_Income <- glm(formula = ObamaWin ~ Income, data = Election08, family = "binomial")
model_HS <- glm(formula = ObamaWin ~ HS, data = Election08, family = "binomial")
model_BA <- glm(formula = ObamaWin ~ BA, data = Election08, family = "binomial")
model_DemRep <- glm(formula = ObamaWin ~ Dem.Rep, data = Election08, family = "binomial")
```

Graphic Representation of the Simple Models

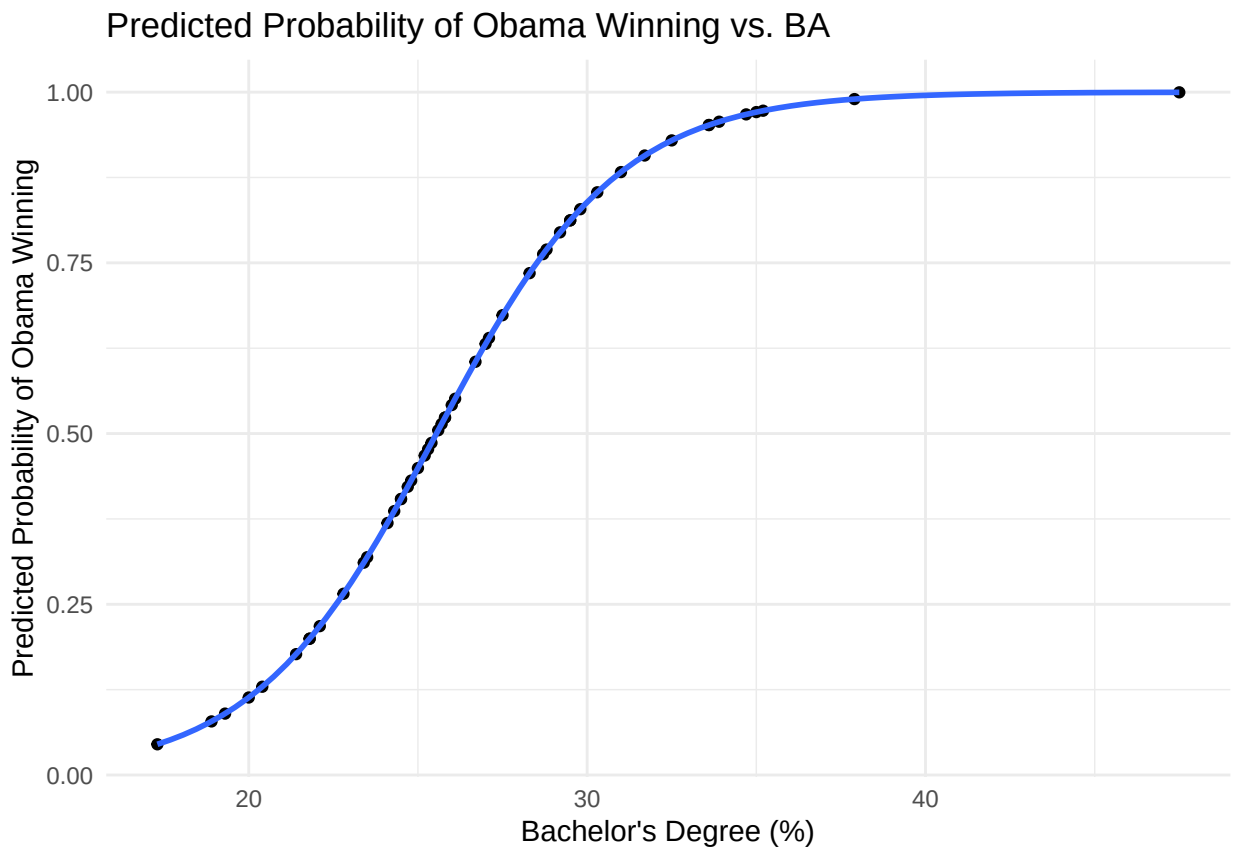
```
Election08$PredictedProb_Income <- predict(model_Income, type = "response")
Election08$PredictedProb_HS <- predict(model_HS, type = "response")
Election08$PredictedProb_BA <- predict(model_BA, type = "response")
Election08$PredictedProb_DemRep <- predict(model_DemRep, type = "response")

ggplot(Election08, aes(x = Income, y = PredictedProb_Income)) +
  geom_point() +

  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +

  # geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(
    title = "Predicted Probability of Obama Winning vs. Income",
    x = "Income (Per Capita)",
    y = "Predicted Probability of Obama Winning"
  ) +
  theme_minimal()
```



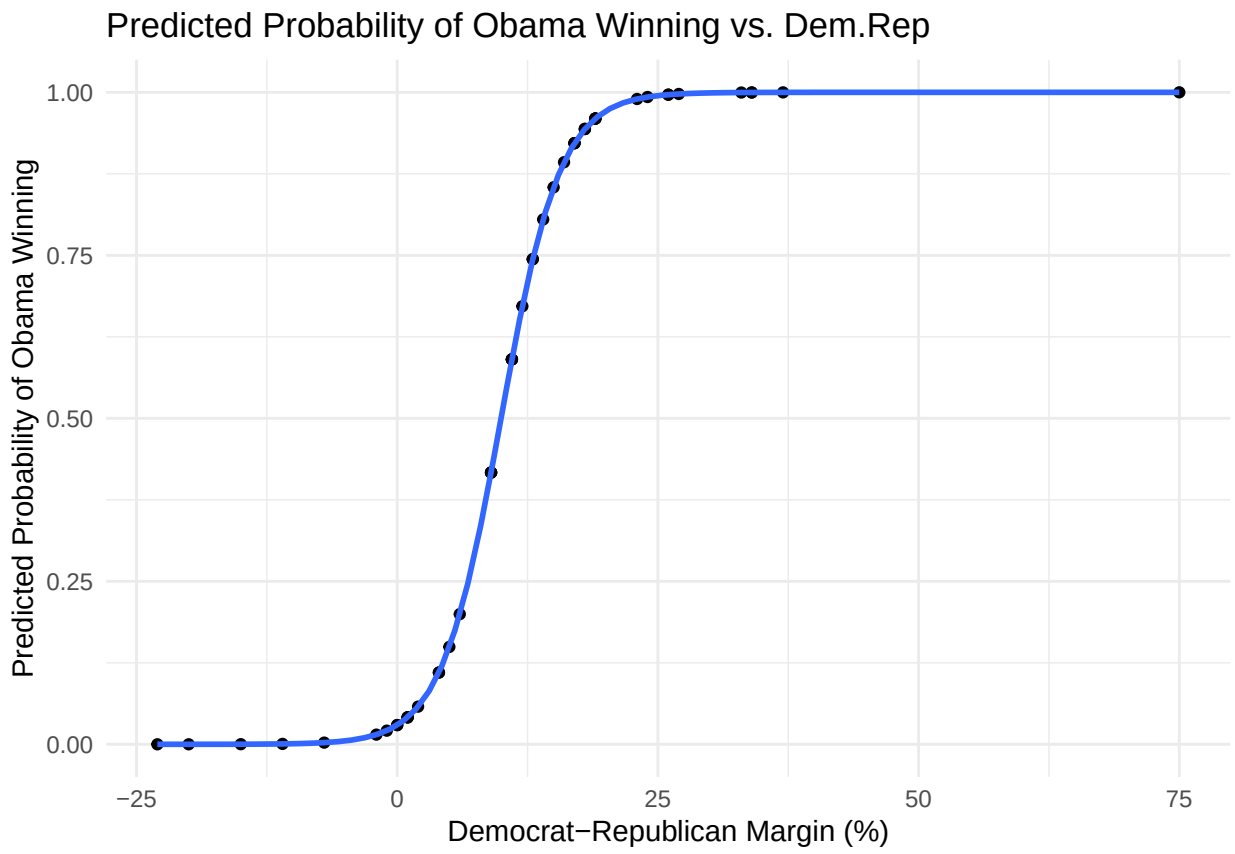

```r
ggplot(Election08, aes(x = HS, y = PredictedProb_HS)) +
  geom_point() +

  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +

  # geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(
    title = "Predicted Probability of Obama Winning vs. HS",
    x = "High School Education (%)",
    y = "Predicted Probability of Obama Winning"
  ) +
  theme_minimal()
```




```
ggplot(Election08, aes(x = BA, y = PredictedProb_BA)) +
  geom_point() +

  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  # geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(
    title = "Predicted Probability of Obama Winning vs. BA",
    x = "Bachelor's Degree (%)",
    y = "Predicted Probability of Obama Winning"
  ) +
  theme_minimal()
```




```r
ggplot(Election08, aes(x = Dem.Rep, y = PredictedProb_DemRep)) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  # geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(
    title = "Predicted Probability of Obama Winning vs. Dem.Rep",
    x = "Democrat-Republican Margin (%)",
    y = "Predicted Probability of Obama Winning"
  ) +
  theme_minimal()
```



Summary of the Simple Logistic Regression Models

```
summary_Income <- summary(model_Income)
summary_HS <- summary(model_HS)
summary_BA <- summary(model_BA)
summary_DemRep <- summary(model_DemRep)

print(summary_Income)
```

```
##
## Call:
## glm(formula = ObamaWin ~ Income, family = "binomial", data = Election08)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.243e+01  3.752e+00  -3.311 0.000928 ***
## Income       3.494e-04  1.050e-04   3.328 0.000874 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.737  on 50  degrees of freedom
```

```
## Residual deviance: 48.867  on 49  degrees of freedom
## AIC: 52.867
##
## Number of Fisher Scoring iterations: 5
```

```
print(summary_HS)
```

```
##
## Call:
## glm(formula = ObamaWin ~ HS, family = "binomial", data = Election08)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -13.63514    7.20057  -1.894   0.0583 .
## HS            0.16195    0.08381   1.932   0.0533 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.737  on 50  degrees of freedom
## Residual deviance: 65.741  on 49  degrees of freedom
## AIC: 69.741
##
## Number of Fisher Scoring iterations: 4
```

```
print(summary_BA)
```

```
##
## Call:
## glm(formula = ObamaWin ~ BA, family = "binomial", data = Election08)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.4667     2.9656  -3.192  0.00141 **
## BA            0.3706     0.1143   3.243  0.00118 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.737  on 50  degrees of freedom
## Residual deviance: 49.689  on 49  degrees of freedom
## AIC: 53.689
##
## Number of Fisher Scoring iterations: 5
```

```
print(summary_DemRep)
```

```
##
## Call:
## glm(formula = ObamaWin ~ Dem.Rep, family = "binomial", data = Election08)
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.4931     1.2290  -2.842 0.004480 **
## Dem.Rep       0.3508     0.1054   3.328 0.000875 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.737  on 50  degrees of freedom
## Residual deviance: 27.167  on 49  degrees of freedom
## AIC: 31.167
##
## Number of Fisher Scoring iterations: 7
```

Based on the summary of the simple logistic models, the AIC values are identified as follows:

1. AIC for ObamaWin ~ Income: 52.867
2. AIC for ObamaWin ~ HS: 69.741
3. AIC for ObamaWin ~ BA: 53.689
4. AIC for ObamaWin ~ Dem.Rep: 31.167

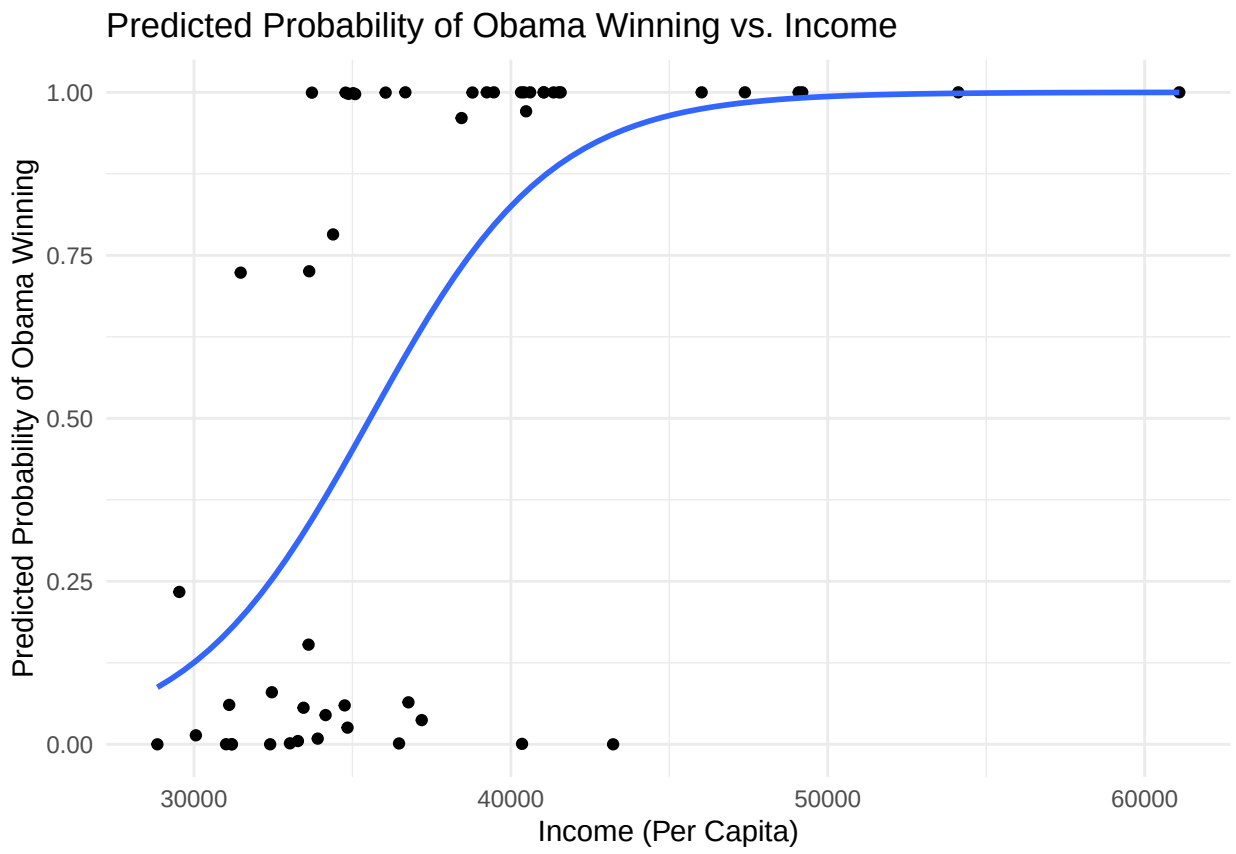
Therefore, among the independent variables, Dem.Rep (political leaning) is the strongest predictor of election outcomes, as evidenced by its substantially lower AIC value (31.167) compared to the other models. On the other hand, socioeconomic variables (Income, HS, and BA) are weaker predictors, with HS being the least effective (AIC = 69.741).

Multiple Logistic Regression Model Construction

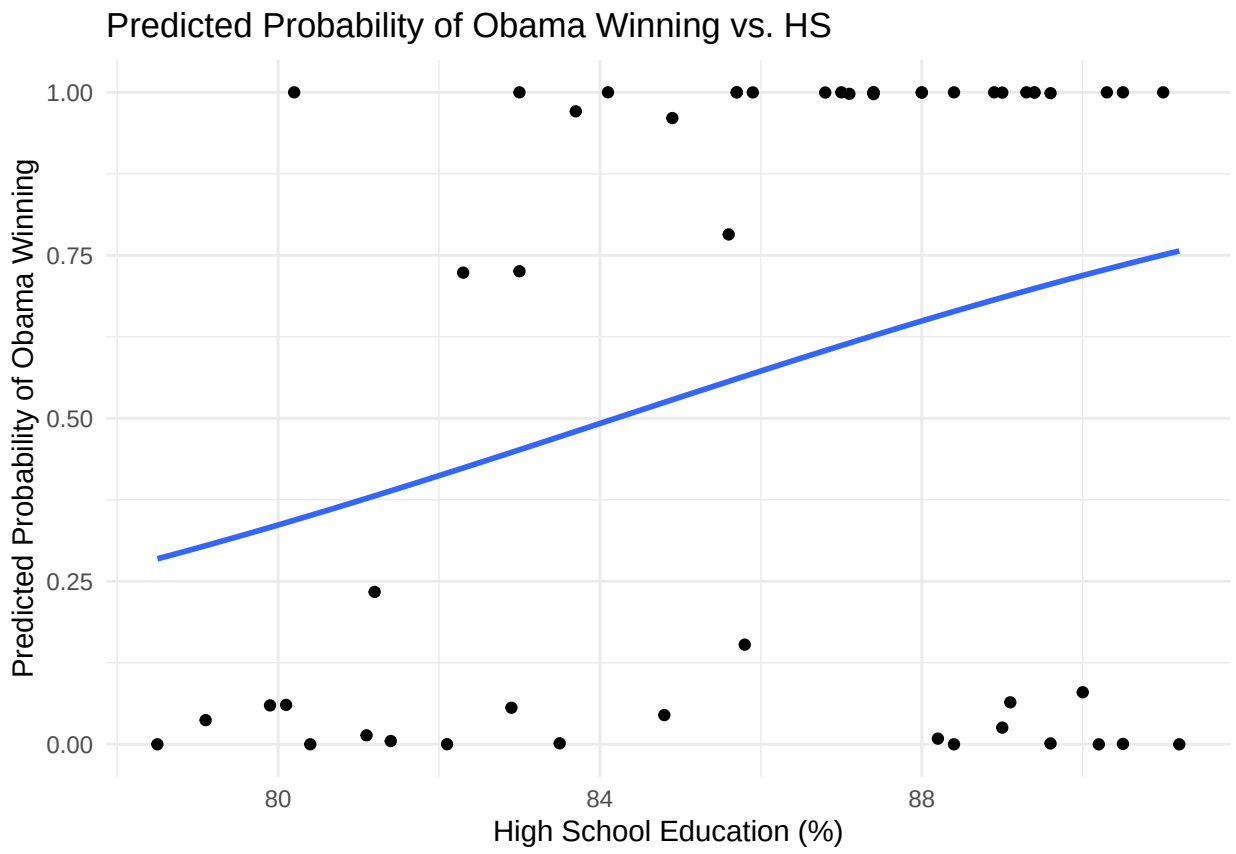
```r
final_model <- glm(formula = ObamaWin ~ Income + HS + BA + Dem.Rep, data = Election08, family = "binomia
```

Graphic Representation of the Final Model

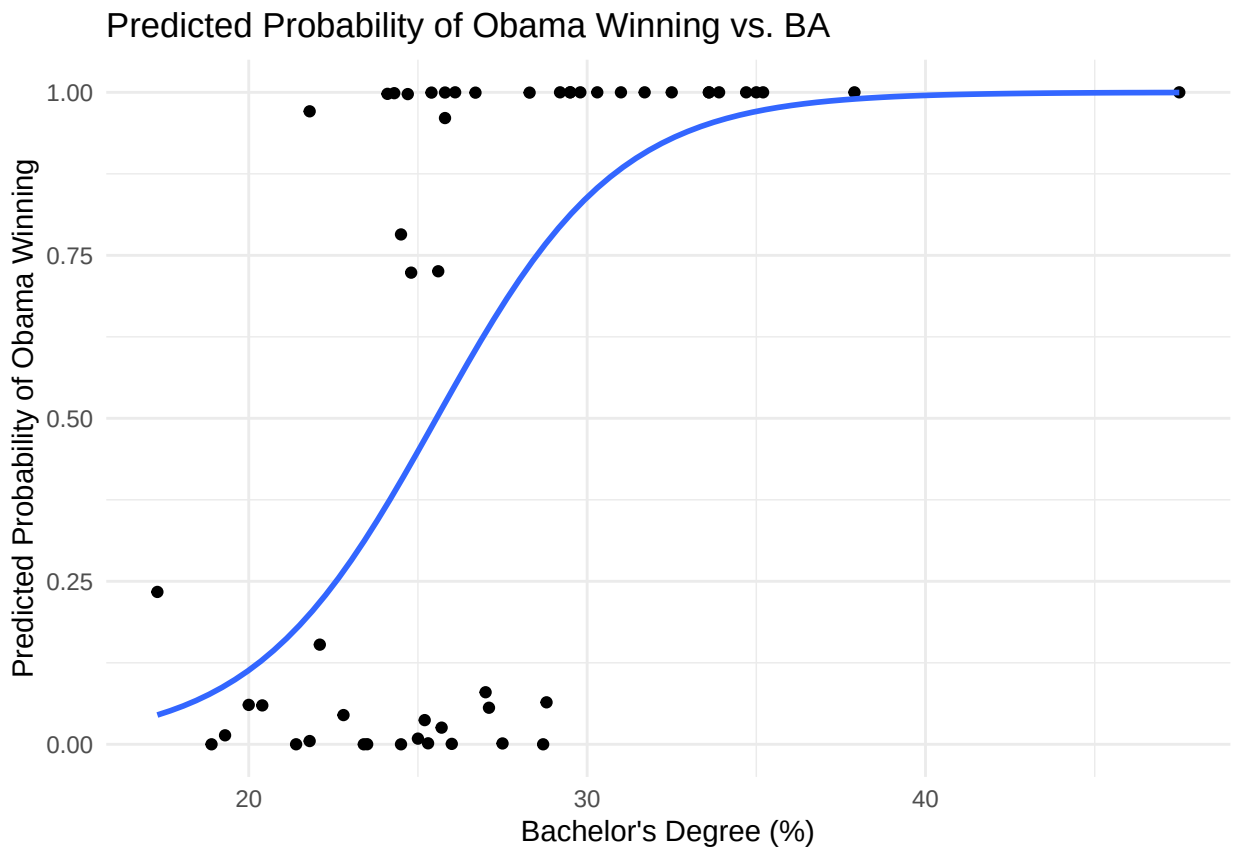
```r
Election08$PredictedProb <- predict(final_model, type = "response")

ggplot(Election08, aes(x = Income, y = PredictedProb)) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  # geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(
    title = "Predicted Probability of Obama Winning vs. Income",
    x = "Income (Per Capita)",
    y = "Predicted Probability of Obama Winning"
  ) +
  theme_minimal()
```



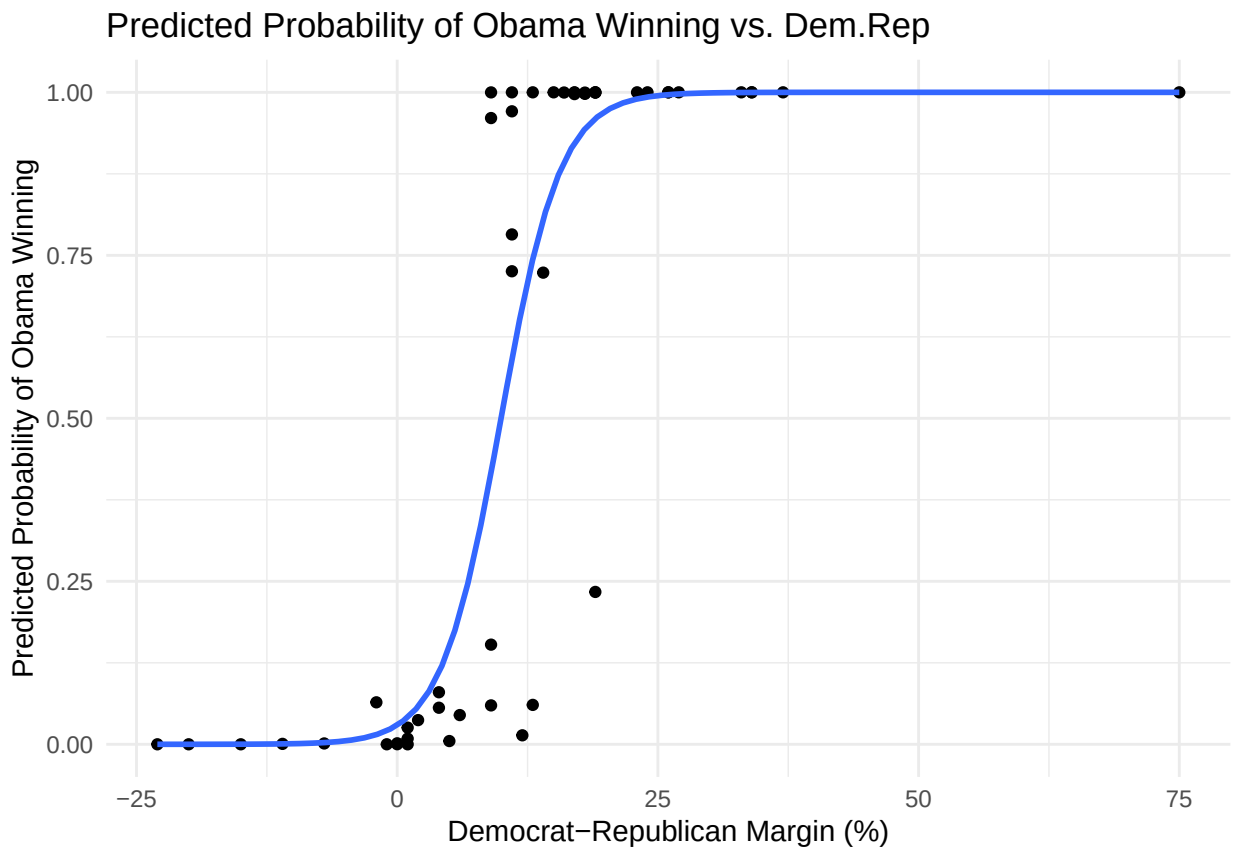

```r
ggplot(Election08, aes(x = HS, y = PredictedProb)) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  # geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(
    title = "Predicted Probability of Obama Winning vs. HS",
    x = "High School Education (%)",
    y = "Predicted Probability of Obama Winning"
  ) +
  theme_minimal()
```




```
ggplot(Election08, aes(x = BA, y = PredictedProb)) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  # geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(
    title = "Predicted Probability of Obama Winning vs. BA",
    x = "Bachelor's Degree (%)",
    y = "Predicted Probability of Obama Winning"
  ) +
  theme_minimal()
```




```
ggplot(Election08, aes(x = Dem.Rep, y = PredictedProb)) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  # geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(
    title = "Predicted Probability of Obama Winning vs. Dem.Rep",
    x = "Democrat-Republican Margin (%)",
    y = "Predicted Probability of Obama Winning"
  ) +
  theme_minimal()
```



Summary of the Multiple Logistic Regression Models

```
summary(final_model)
```

```
##
## Call:
## glm(formula = ObamaWin ~ Income + HS + BA + Dem.Rep, family = "binomial",
##     data = Election08)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.361e+01  3.773e+01  -1.421   0.1554
## Income       6.445e-04  4.828e-04   1.335   0.1819
## HS           1.514e-01  3.894e-01   0.389   0.6973
## BA           5.214e-01  3.947e-01   1.321   0.1865
## Dem.Rep      6.353e-01  2.726e-01   2.331   0.0198 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.7372  on 50  degrees of freedom
## Residual deviance:  9.7252  on 46  degrees of freedom
```

```
## AIC: 19.725
##
## Number of Fisher Scoring iterations: 10
```

Interpretation of the Estimated Coefficients from the Final Model

The final multiple logistic regression model includes the predictors Income, HS, BA, and Dem.Rep to predict whether Barack Obama won the 2008 U.S. presidential election.

1. Intercept:
    - Estimate: -53.61
    - Interpretation: When all predictor variables (Income, HS, BA, and Dem.Rep) are zero, the log-odds of Obama winning a state are -53.61. However, this value has no practical meaning because all predictors being zero is unrealistic in this context.
    - Significance: With a p-value of 0.1554, the intercept is not statistically significant.
2. Income (Per Capita Income):
    - Estimate: 0.000645
    - Interpretation: For every $1 increase in per capita income, the log-odds of Obama winning increase by 0.000645, holding all other predictors constant. This suggests a very small positive effect of income on the likelihood of Obama winning.
    - Significance: With a p-value of 0.1819, the effect of income is not statistically significant.
3. HS (Percentage of Adults with at Least a High School Education):
    - Estimate: 0.1514
    - Interpretation: For every 1% increase in the percentage of adults with at least a high school education, the log-odds of Obama winning increase by 0.1514, holding all other predictors constant. This indicates a weak positive relationship between high school education rates and the likelihood of Obama winning.
    - Significance: With a p-value of 0.6973, this predictor is not statistically significant.
4. BA (Percentage of Adults with at Least a Bachelor's Degree):
    - Estimate: 0.5214
    - Interpretation: For every 1% increase in the percentage of adults with at least a bachelor's degree, the log-odds of Obama winning increase by 0.5214, holding all other predictors constant. This indicates a moderate positive effect of college education on the likelihood of Obama winning.
    - Significance: With a p-value of 0.1865, this effect is not statistically significant.
5. Dem.Rep (Difference in Percent Democrat and Percent Republican):
    - Estimate: 0.6353
    - Interpretation: For every 1% increase in the Democrat-Republican margin, the log-odds of Obama winning increase by 0.6353, holding all other predictors constant. This demonstrates a strong positive relationship between political leaning and the likelihood of Obama winning.
    - Significance: With a p-value of 0.0198, this predictor is statistically significant at the 5% level, making it the most important variable in the model.

Model Comparison by G-tests

```r
## G-test for Income
G <- 69.737 - 48.867
df <- 1
p_value_g_test <- pchisq(G, df, lower.tail = FALSE)
print(p_value_g_test)
```

```
## [1] 4.915329e-06
```

```r
## G-test for HS
G <- 69.737 - 65.741
df <- 1
p_value_g_test <- pchisq(G, df, lower.tail = FALSE)
print(p_value_g_test)
```

```
## [1] 0.04560838
```

```r
## G-test for BA
G <- 69.737 - 49.689
df <- 1
p_value_g_test <- pchisq(G, df, lower.tail = FALSE)
print(p_value_g_test)
```

```
## [1] 7.552247e-06
```

```r
## G-test for DemRep
G <- 69.737 - 27.167
df <- 1
p_value_g_test <- pchisq(G, df, lower.tail = FALSE)
print(p_value_g_test)
```

```
## [1] 6.819737e-11
```

```r
## G-test for Final Model
G <- 69.737 - 9.7252
df <- 4
p_value_g_test <- pchisq(G, df, lower.tail = FALSE)
print(p_value_g_test)
```

```
## [1] 2.884347e-12
```

Interpretation of G-Test Results

1. G-Test for Income:

- Test Statistic (G): G = 69.737 - 48.867 = 20.87
- Degrees of Freedom (df): 1
- p-value: 4.92*10^-6

- Interpretation: The small p-value indicates that adding Income to the null model significantly improves the model's fit. This suggests that Income is a meaningful predictor of election outcomes when considered on its own.

2. G-Test for HS (High School Education):

- Test Statistic (G): G = 69.737 - 65.741 = 3.996
- Degrees of Freedom (df): 1
- p-value: 0.0456
- Interpretation: The p-value indicates that HS marginally improves the model's fit over the null model. However, its effect is weaker compared to other predictors, and its impact may not be substantial in the presence of other variables.

3. G-Test for BA (Bachelor's Education):

- Test Statistic (G): G = 69.737 - 49.689 = 20.048
- Degrees of Freedom (df): 1
- p-value: 7.55*10^-6
- Interpretation: The small p-value suggests that BA significantly improves the model's fit over the null model, indicating it is a meaningful predictor of election outcomes when considered individually.

4. G-Test for Dem.Rep (Democrat-Republican Margin):

- Test Statistic (G): G = 69.737 - 27.167 = 42.57
- Degrees of Freedom (df): 1
- p-value: 6.82*10^-11
- Interpretation: The extremely small p-value indicates that Dem.Rep significantly improves the model's fit over the null model. This confirms that political leaning is the strongest individual predictor of the election outcomes in 2008.

5. G-Test for Final Model:

- Test Statistic (G): G = 69.737 - 9.7252 = 60.01
- Degrees of Freedom (df): 4
- p-value: 2.37*10^-12
- Interpretation: The final model, which includes Income, HS, BA, and Dem.Rep, significantly improves the fit compared to the null model. The small p-value shows that the combination of all predictors collectively explains the variability in the election outcomes much better than the null model.

Statistical Inference

```
conf_intervals <- confint(final_model)
print(conf_intervals)
```

```
##                     2.5 %      97.5 %
## (Intercept) -1.743312e+02 2.160858247
## Income       3.567895e-05 0.002271797
## HS          -6.186328e-01 1.222498553
## BA          -1.817063e-01 1.493866249
## Dem.Rep      2.655156e-01 1.478897185
```

Results & Interpretation

1. Simple Logistic Regression Models:
    - Each predictor (Income, HS, BA, and Dem.Rep) was independently analyzed for its relationship with the likelihood of Barack Obama winning a state.
    - Among these, Dem.Rep had the strongest and most statistically significant effect, with a coefficient of 0.3508 ($p < 0.001$). This indicates that for every 1% increase in the Democrat-Republican margin, the log-odds of Obama winning the state increase by 0.3508.
    - Income and BA were also significant predictors ($p < 0.01$), with higher income and higher college education rates associated with an increased likelihood of Obama winning.
    - HS was marginally significant, showing a weaker effect compared to other predictors.
2. Multiple Logistic Regression Model:
    - When combining all predictors (Income, HS, BA, and Dem.Rep) into a single model, only Dem.Rep remained significant ($p = 0.0198$).
    - This result suggests that Dem.Rep is the most influential factor when controlling for the other variables.
    - The combined model had a much lower residual deviance (9.7252) and AIC (19.725), indicating better overall fit compared to individual models.
3. Model Comparison:
    - The multiple logistic regression model performed better than the simple models.
    - The G-test confirmed that the final model explains significantly more variability in the election outcomes compared to the null model ($p < 0.001$).
4. Confidence Interval
    - Income:
        - 95% CI: (0.0000357, 0.0022718)
        - Interpretation: For every 1-unit increase in per capita income (in dollars), the log-odds of Obama winning increases by a value between 0.0000357 and 0.0022718, holding all other predictors constant.
        - Since the interval does not include 0, Income is likely to have a statistically significant, positive effect on the likelihood of Obama winning.
    - HS (Percentage of High School Graduates):
        - 95% CI: (-0.6186, 1.2225)
        - Interpretation: For every 1% increase in high school education attainment, the log-odds of Obama winning changes between -0.6186 and 1.2225, holding all other predictors constant.
        - The interval includes 0, suggesting that the effect of high school education on the likelihood of Obama winning is not statistically significant in this model.
    - BA (Percentage of Bachelor's Degree Holders):
        - 95% CI: (-0.1817, 1.4939)
        - Interpretation: For every 1% increase in college education attainment, the log-odds of Obama winning changes between -0.1817 and 1.4939, holding all other predictors constant.
        - The interval includes 0, indicating that the effect of college education on the likelihood of Obama winning is not statistically significant.
    - Dem.Rep (Difference in Percent Democrat and Percent Republican):
        - 95% CI: (0.2655, 1.4789)
        - Interpretation: For every 1% increase in the difference between Democratic and Republican support, the log-odds of Obama winning increases by a value between 0.2655 and 1.4789, holding all other predictors constant.
        - Since the interval does not include 0, Dem.Rep is statistically significant, and its positive effect highlights it as the most influential factor in the model.

Key Findings

1. The analysis confirms that political leaning (Dem.Rep) is the strongest predictor of the 2008 presidential election results. This aligns with the hypothesis that states with a higher Democrat-Republican margin were more likely to vote for Barack Obama
2. While socioeconomic factors (Income, HS, and BA) are individually significant in some cases, their effects diminish when combined with Dem.Rep, likely due to overlapping explanatory power.
3. The multiple logistic regression model outperforms the simple models by capturing the joint effects of all predictors, making it the most effective approach for predicting state-level election outcomes.

Answers to the Research Questions

1. Dem.Rep explains the majority of the variation in the election outcomes.
2. The multiple logistic regression model provides the best predictive performance due to its ability to account for interactions and combined effects of predictors.

Suggestions for Future Research

1. Inclusion of Additional Variables: Future studies could incorporate other influential factors such as voter turnout rates, urban versus rural demographics, or historical voting patterns to gain a more comprehensive understanding of election outcomes.
2. Temporal Expansion: Analyzing data from other presidential elections could help validate whether Dem.Rep remains the most influential predictor across different election years.
3. Regional Analysis: Conducting regional studies (for example, Midwest vs. South) could provide insights into whether the influence of political leaning varies geographically.
4. Interaction Effects: Future research could explore interactions between predictors, such as how income levels might moderate the effect of education on election outcomes.

Implications

1. Political Strategy: The strong influence of Dem.Rep highlights the importance of targeted campaigning in states with close Democrat-Republican margins. Understanding political leanings at the state level can guide resource allocation and strategy.
2. Policy Prioritization: Socioeconomic factors like income and education, while secondary, still contribute to election outcomes. Policymakers may consider how addressing these issues aligns with voter preferences.
3. Model Application: The multiple logistic regression model's effectiveness suggests its utility for other binary prediction tasks, such as forecasting election results or voter turnout based on demographic factors.

Limitations

1. Restricted Dataset: The analysis is based on a single election year (2008), limiting the generalizability of findings to other elections.
2. Simplistic Model Assumptions: The logistic regression models assume linear relationships between predictors and the log odds of winning, which may oversimplify complex voter behaviors.
3. Data Availability: Some variables, like detailed demographic breakdowns, were unavailable and could provide a richer analysis.

4. Binary Outcome: The analysis simplifies the election outcome to a binary variable, ignoring nuances such as the margin of victory.

Conclusion

This study identifies Dem.Rep, the difference between the percentage of Democrats and Republicans in a state, as the most influential factor in predicting the 2008 U.S. presidential election outcome. While socioeconomic variables such as income and education levels play a role, their effects are secondary to political leaning. Moreover, the multiple logistic regression model proves to be the most effective for predicting election outcomes, as it accounts for the combined influence of all predictors. These findings underscore the importance of political alignment and multivariable approaches in understanding electoral behavior, providing valuable insights for future analyses and campaign strategies.