10.12 Final Project

Sefu Boisrond

2024-05-27

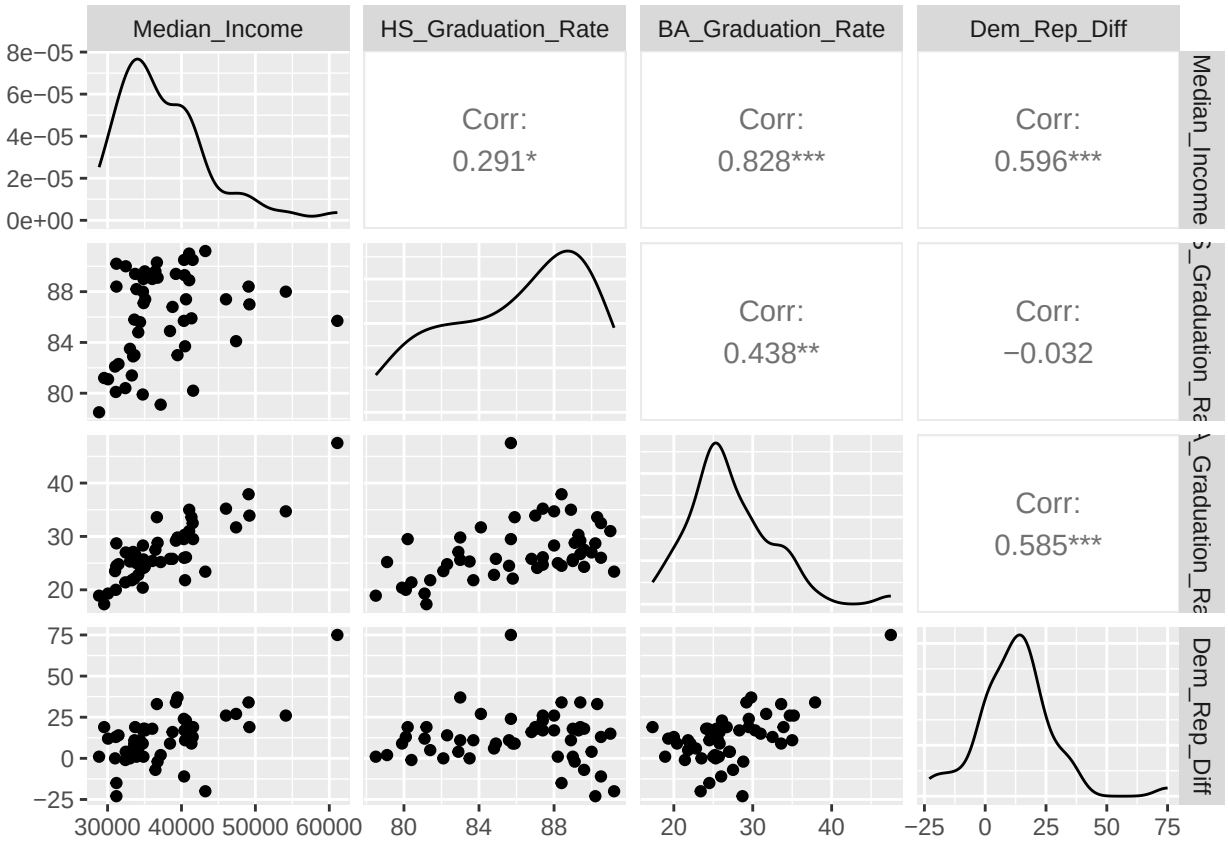
Dataset: 2008 U.S. Presidential Election

Research Question: "How do voter demographics (age, education level, and income) influence the voting preference for the Democratic or Republican candidate in the 2008 U.S. Presidential Election?"

This question examines the relationship between demographic factors and voting preferences, allowing us to analyze how different segments of the population voted.

Visualization

```
# Visualize the relationships between demographics and voting preference
# Scatter plot matrix
ggpairs(election_data, columns = c("Median_Income", "HS_Graduation_Rate", "BA_Graduation_Rate", "Dem_Rep
```

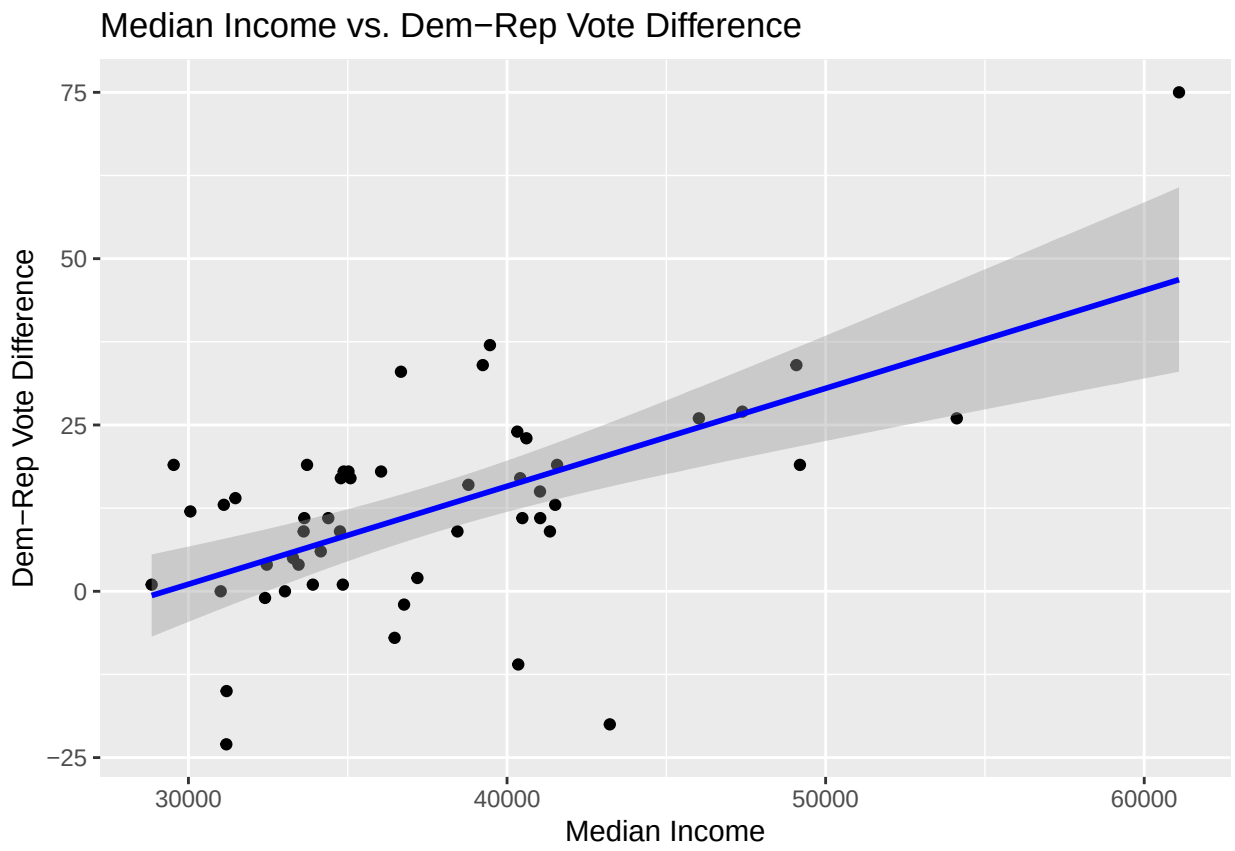


```r
# Linear Regression to predict voting preference based on demographics
model <- lm(Dem_Rep_Diff ~ Median_Income + HS_Graduation_Rate + BA_Graduation_Rate, data = election_data)
summary(model)
```

```
##
## Call:
## lm(formula = Dem_Rep_Diff ~ Median_Income + HS_Graduation_Rate +
##     BA_Graduation_Rate, data = election_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.0263  -8.8250  -0.4185   7.4235  22.5475
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        71.8054473 42.7007066   1.682  0.09928 .
## Median_Income       0.0006887  0.0004727   1.457  0.15175
## HS_Graduation_Rate -1.4493119  0.5215142  -2.779  0.00781 **
## BA_Graduation_Rate  1.4448764  0.5887584   2.454  0.01788 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.9 on 47 degrees of freedom
## Multiple R-squared:  0.4691, Adjusted R-squared:  0.4352
## F-statistic: 13.84 on 3 and 47 DF,  p-value: 1.345e-06
```

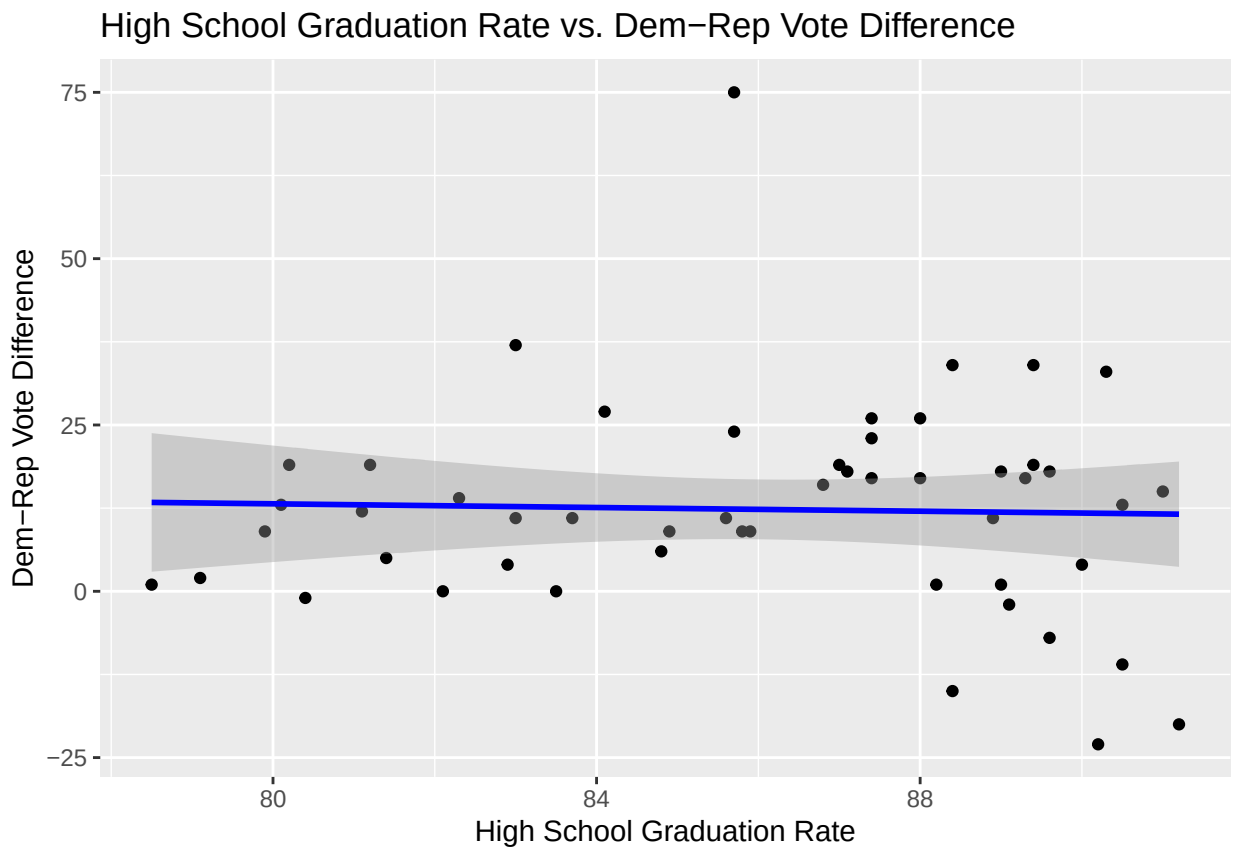
```r
# Check for non-linearity and other issues
#plot(model)

# Visualize the relationship with regression lines
ggplot(election_data, aes(x = Median_Income, y = Dem_Rep_Diff)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Median Income vs. Dem-Rep Vote Difference", x = "Median Income", y = "Dem-Rep Vote Diff
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



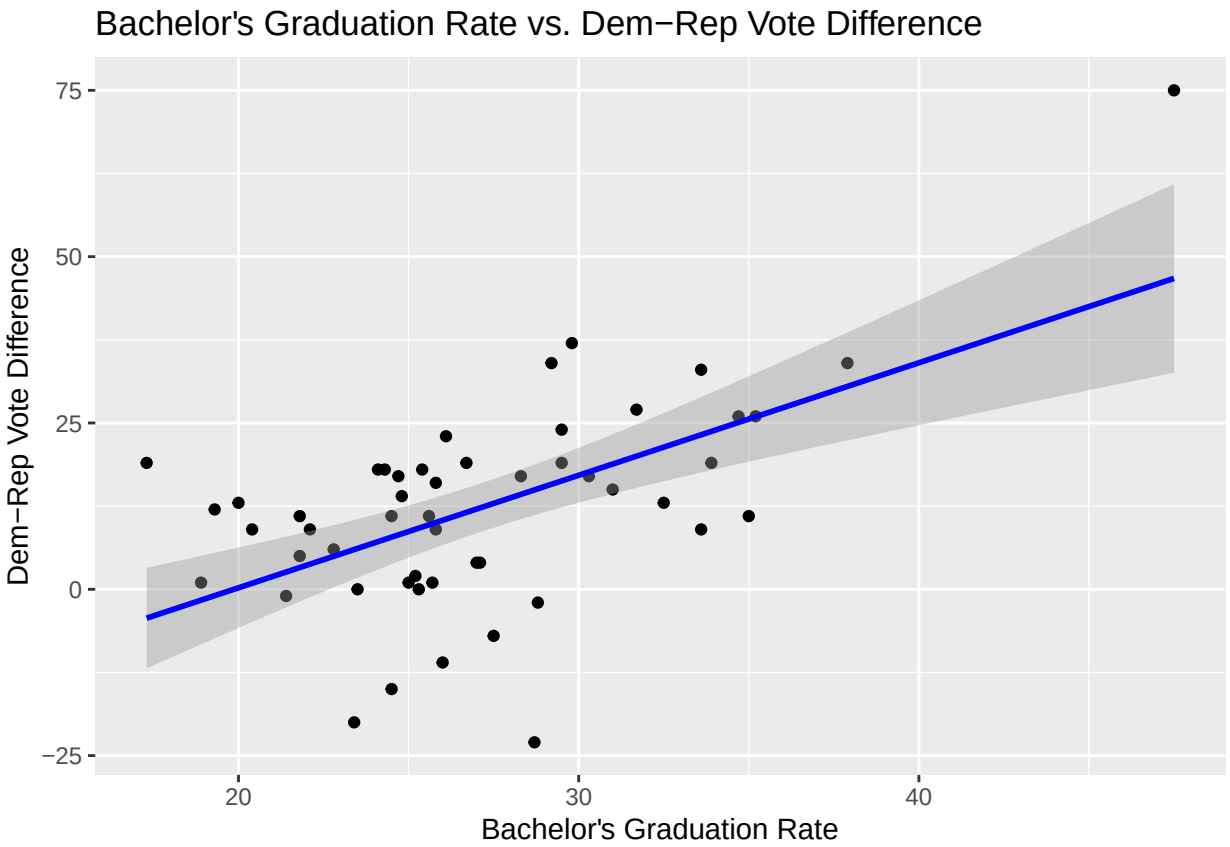

```
ggplot(election_data, aes(x = HS_Graduation_Rate, y = Dem_Rep_Diff)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "High School Graduation Rate vs. Dem-Rep Vote Difference", x = "High School Graduation Ra
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(election_data, aes(x = BA_Graduation_Rate, y = Dem_Rep_Diff)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Bachelor's Graduation Rate vs. Dem-Rep Vote Difference", x = "Bachelor's Graduation Rat
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



1. Dataset Description The dataset used is the "Cleaned_Election_Data.csv," which contains information on voter demographics and voting outcomes from the 2008 U.S. Presidential Election. This dataset was chosen because it provides comprehensive information on demographic factors and their potential influence on voting preferences, making it suitable for studying the relationship between these variables and voting outcomes.Each line in the dataset represents data for a specific state in the U.S. during the 2008 Presidential Election.

Columns and Meanings:

State: The name of the state. State_Abbr: The abbreviation of the state. Median_Income: The median household income in the state. HS_Graduation_Rate: The high school graduation rate in the state. BA_Graduation_Rate: The bachelor's degree graduation rate in the state. Dem_Rep_Diff: The difference in votes between Democratic and Republican candidates. Obama_Win: Indicates whether Obama won in the state (1 if yes, 0 if no).

2. Research Question Research Question: "How do voter demographics (median income, high school graduation rate, and bachelor's degree graduation rate) influence the voting preference for the Democratic or Republican candidate in the 2008 U.S. Presidential Election?"

Interest in Relationship: Understanding the influence of voter demographics on voting preferences can provide insights into the factors driving electoral outcomes, which is important for political strategists, sociologists, and policymakers.Variables such as state names and abbreviations were excluded from the analysis because they do not directly contribute to understanding the demographic influences on voting preferences.

3. Method Selection Method: Multivariate linear regression was chosen to analyze the relationship between voter demographics and voting preference.

Strengths of Method:

- Interpretability: Linear regression provides clear coefficients that quantify the relationship between predictors and the outcome variable.
- Statistical Significance: The method allows for hypothesis testing to determine the significance of each predictor.
- Control for Confounding: It can control for multiple predictors simultaneously, isolating the effect of each variable.

4. Plots

5. Regression Equation

Dem_Rep_Diff = 0 + 1 * Median_Income + 2 * HS_Graduation_Rate + 3 * BA_Graduation_Rate Where: • 0 is the intercept. • 1, 2, and 3 are the coefficients for Median Income, High School Graduation Rate, and Bachelor's Degree Graduation Rate, respectively.

6.

Coefficients and Interpretation

Intercept (0):

Estimate: 71.8054 Standard Error: 42.7007 t value: 1.682 p-value: 0.09928 Interpretation: The intercept represents the expected difference in votes between Democratic and Republican candidates when all predictors are zero. Although not practically meaningful (since demographics can't be zero), it provides the baseline level of the outcome variable.

Median_Income (1):

Estimate: 0.0007 Standard Error: 0.0005 t value: 1.457 p-value: 0.15175 Interpretation: A $1 increase in median income is associated with an increase of 0.0007 in the difference in votes between Democratic and Republican candidates. However, this relationship is not statistically significant ($p > 0.05$), suggesting that median income may not have a strong influence on voting preference in this model.

HS_Graduation_Rate (2):

Estimate: -1.4493 Standard Error: 0.5215 t value: -2.779 p-value: 0.00781 Interpretation: A 1% increase in the high school graduation rate is associated with a decrease of 1.4493 in the difference in votes between Democratic and Republican candidates. This relationship is statistically significant ($p < 0.01$), indicating that higher high school graduation rates are associated with lower Democratic vote margins relative to Republican votes.

BA_Graduation_Rate (3):

Estimate: 1.4449 Standard Error: 0.5888 t value: 2.454 p-value: 0.01788 Interpretation: A 1% increase in the bachelor's degree graduation rate is associated with an increase of 1.4449 in the difference in votes between Democratic and Republican candidates. This relationship is statistically significant ($p < 0.05$), indicating that higher rates of bachelor's degree attainment are associated with higher Democratic vote margins relative to Republican votes.

7. Analytical Weaknesses of Multivariate Linear Regression

- Linearity Assumption:

  Issue: Linear regression assumes a linear relationship between the predictors and the response variable. If the true relationship is nonlinear, the model may not fit the data well, leading to inaccurate predictions. Implication: When the linearity assumption is violated, it can result in biased estimates and misleading conclusions.

- Multicollinearity:

  Issue: Multicollinearity occurs when predictor variables are highly correlated with each other. This can inflate the standard errors of the coefficients, making it difficult to determine the individual effect of each predictor. Implication: Multicollinearity can make the model unstable and the results difficult to interpret, as small changes in the data can lead to large changes in the estimated coefficients.

  Implications

- Scientific Perspective:

  Bias and Misinterpretation: Scientific conclusions based on biased or inaccurate models can misguide future research. Incorrect inferences about the relationship between demographics and voting behavior can lead to flawed theories and models. Generalizability: If the model's assumptions are not met, the findings may not generalize to other contexts or populations, limiting the applicability of the research.

- Moral Perspective:

  Policy Decisions: Policymakers often rely on scientific research to make decisions. If the analysis is flawed, it could lead to policies that are not effective or fair. For example, misinterpreting the influence of education on voting behavior might result in policies that do not address the actual needs of the population. Fair Representation: Incorrect conclusions about voter behavior can result in certain demographic groups being unfairly represented or targeted, perpetuating inequality and bias.

- Societal Perspective:

  Public Trust: Inaccurate or biased research can erode public trust in scientific studies and institutions. If the public perceives that research is flawed or biased, they may be less likely to support or believe in scientific findings. Electoral Strategy: Misleading conclusions about voting behavior can affect electoral strategies, potentially leading to ineffective campaigning and resource allocation. This can distort the democratic process, as parties may not engage with or represent the true interests of the electorate.

Addressing Weaknesses

To mitigate these weaknesses, consider the following approaches:

Nonlinear Models: If a linear relationship is not appropriate, consider using nonlinear regression models or transformation of variables. Multicollinearity Checks: Use techniques such as variance inflation factor (VIF) to detect multicollinearity and consider removing or combining correlated predictors.