# Undercounted votes in Georgia in the 2000 US presidential election

Herman Githinji Mwangi *

Policy analysis blog

(2795 words)

May 2023

*githinjiherman@gmail.com

Abstract

The study aims to determine a thorough overview on Georgia's 2000 presidential elections. The high residual voting rate of 3.5 percent, which is the second-worst in the nation, is what makes this intriguing. What factors led to the lost votes in Georgia's 2000 presidential elections? It also seeks to explain the origins of this inaccuracy. By identifying the elements that influenced the percentage of residual votes in Georgia's elections, this study adds to the body of knowledge on the voting problems. The residual vote ratio in the US election system is a significant concern, according to prior study. In order to identify random variables, the study used data analysis and pertinent linear models, including estimate, analysis, and model selection. Each instance in the research represents a county in Georgia, and there are 159 samples with 10 variables in total. To learn in-depth details about the elements that affected the residual voting rate, the case study and interview techniques were selected. The study's findings can be applied to advance voting technology and lower the residual vote rate in next elections.

# Contents

# 1 Introduction

President Bill Clinton's two terms in office came to an end with the historic 2000 US presidential election. The election was a hotly contested struggle between Republican George W. Bush, the Texas governor at the time, and Democrat Al Gore, the former vice president. The election, which was held on November 7, 2000, resulted in a court dispute that dragged on for several weeks without a conclusion. In the end, the Supreme Court awarded Bush a narrow victory over Kerry in Florida and the presidential race. If the uncounted votes from Georgia had been taken into account, the election's result may have been different. Georgia reported 2.60 million votes for the US president out of a total vote count of 2.69 million. Georgia had the nation's second-worst "residual vote rate," a measurement of "lost votes," at 3.5 percent, after only Illinois. This raises concerns regarding the accuracy of voting equipment and the possible influence of invalid ballots on election outcomes. This study tries to investigate the reasons for Georgia's high residual vote rate in the 2000 presidential election. To identify random variables, the study will make use of data analytics and pertinent linear models, including estimate, analysis, and model selection. The study will concentrate on 159 samples across 10 variables, with each example denoting a Georgian county. The study attempts to close a knowledge vacuum in the area and advance our knowledge of how uncounted ballots affect election outcomes. The study's conclusions will have broader ramifications and may be put to use in the real world to increase the accuracy and dependability of voting technologies.

# 2 Background to the study

Vote decline, or the discrepancy between the number of legitimate votes counted and the number of votes cast, is a major problem in the US. This

reduction can be attributed to a number of factors, such as voters opting not to support any presidential candidates, people casting multiple ballots that are rejected, voters misusing the voting machines or the machines failing to record their selections. One specific worry is that improper use of the voting machines might lower the number of valid votes, especially in places that cannot afford the finest devices. State electoral bodies are therefore interested in analyzing how the voting machinery affected the undercounted ballots. In the Georgia state during the 2000 US presidential election, this research tries to look at the signs of a vote reduction. The state had the second-worst residual vote rate in the nation, trailing only Illinois with 3.5 percent, a measure of invalid votes used to assess the accuracy of voting equipment. In order to determine the reasons for this high proportion of undercounted votes, the study examines 159 counties in Georgia and analyzes 10 different factors. To identify the random variables and account for pertinent elements like the size of the county, the researchers use data analytics and pertinent linear models, including estimate, analysis, and model selection. By doing this, this study hopes to aid in improving knowledge of the effects of voting machinery on the validity of votes and provide insights into potential solutions to address the issue of vote decline in future elections.

# 3 Study Objectives

This study's goal is to determine how voting technology affects Georgia's voter undercount. The purpose of the study is to assess the discrepancy between the total number of votes cast and the total number of legitimate votes counted and to determine the causes of the reduction. While a number of things, such as voters not voting for a specific candidate, voting more than once, or failing to follow instructions when using electoral equipment, can

result in undercounts, the study aims to investigate how voting machines affect the validity of the votes.

# 4 Brief description of the data

The data used in this analysis is shown in Table 1.A data frame with 159 observations on the following 10 variables. Each case represents a county in Georgia.

Table 1: Data used in the analysis

| **Variable** | **Description** |
|---|---|
| equip | The voting equipment used:<br>LEVER, OS-CC (optical, central count),<br>OS-PC (optical, precinct count),<br>PAPER and PUNCH |
| econ | Economic status of county: middle, poor, rich |
| perAA | Percent of African Americans in county |
| rural | Indicator of whether county is rural or urban |
| atlanta | Indicator of whether county is in Atlanta or not: notAtlanta |
| gore | Number of votes for Gore |
| bush | Number of votes for Bush |
| other | Number of votes for other candidates |
| votes | Number of votes |
| ballots | Number of ballots |

# 5 Data analysis and interpretation

An essential initial step in any statistical study is exploratory and descriptive data analysis. Visually examining the data is necessary to spot trends and probable outliers in this situation. The process of fitting linear models to the data comes next. While diagnostics are used to verify model assumptions like the normality of residuals, constant variance, and independence of errors, estimation is used to determine the values of the model parameters. Finding a model that can be used to generate precise predictions or inferences about

the population and that explains the link between the response variable and the predictor factors in the data is the aim of this approach.

# 6 Descriptive analysis

This method involves summarizing the main features of a dataset, such as central tendency, variability, and distribution, using statistical methods and visual tools like graphs and charts.

## 6.1 Summary Statistics

A summary of the data is shown in the table below:

Summary of Data

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| perAA | 159 | 0.243 | 0.163 | 0.000 | 0.765 |
| gore | 159 | 7,020.314 | 19,317.780 | 249 | 154,509 |
| bush | 159 | 8,929.057 | 18,029.960 | 271 | 140,494 |
| other | 159 | 381.654 | 1,150.975 | 5 | 7,920 |
| votes | 159 | 16,331.020 | 36,623.270 | 832 | 263,211 |
| ballots | 159 | 16,926.500 | 37,865.150 | 881 | 280,975 |

The summary table Table 6.1 provides an overview of the distribution of the variables in the dataset. The variable 'perAA' has a mean of 0.243 and a standard deviation of 0.163, indicating that the percentage of African Americans in each county ranges from 0 percent to 76.5 percent with an average of 24.3 percent. The 'gore' variable has a mean of 7,020.314 and a standard

deviation of 19,317.780, indicating that the number of votes for Al Gore in each county ranges from 249 to 154,509 with an average of 7,020. The 'bush' variable has a mean of 8,929.057 and a standard deviation of 18,029.960, indicating that the number of votes for George W. Bush in each county ranges from 271 to 140,494 with an average of 8,929. The 'other' variable has a mean of 381.654 and a standard deviation of 1,150.975, indicating that the number of votes for other candidates in each county ranges from 5 to 7,920 with an average of 381. The 'votes' variable has a mean of 16,331.020 and a standard deviation of 36,623.270, indicating that the total number of votes cast in each county ranges from 832 to 263,211 with an average of 16,331. Finally, the 'ballots' variable has a mean of 16,926.500 and a standard deviation of 37,865.150, indicating that the total number of ballots cast in each county ranges from 881 to 280,975 with an average of 16,926.5.

## 6.2 Data

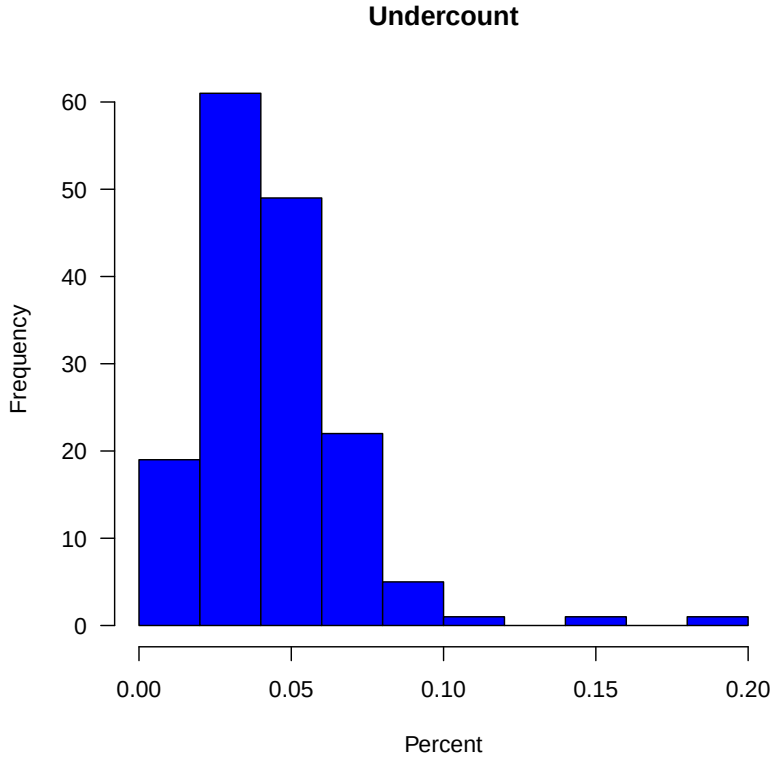
The first few lines of the data to see the variables:

Table: Some values of the Data Frame

| | equip | econ | perAA | rural | atlanta | gore | bush | other | |
|---|---|---|---|---|---|---|---|---|---|
| APPLING | LEVER | poor | 0.182 | rural | notAtlanta | 2093 | 3940 | 66 | |
| ATKINSON | LEVER | poor | 0.230 | rural | notAtlanta | 821 | 1228 | 22 | |
| BACON | LEVER | poor | 0.131 | rural | notAtlanta | 956 | 2010 | 29 | |
| BAKER | OS-CC | poor | 0.476 | rural | notAtlanta | 893 | 615 | 11 | |
| BALDWIN | LEVER | middle | 0.359 | rural | notAtlanta | 5893 | 6041 | 192 | |
| BANKS | LEVER | middle | 0.024 | rural | notAtlanta | 1220 | 3202 | 111 | |

## 6.3 Distribution of Percent Undercount

The distribution of the undercounted votes as a percentage for every county in Georgia and its frequency are shown and relayed on a histogram. The area is proportional to the frequency of a variable and whose width is equal to the class interval.



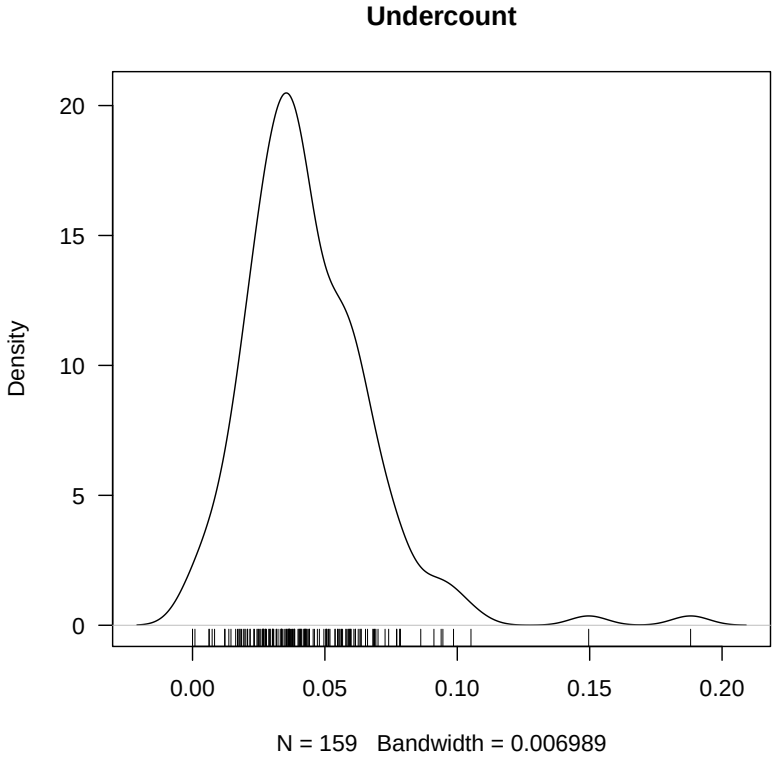

The histogram is unimodal and skewed to the right as it has a peak to the left of center. The mean of the Undercounted votes per county with the highest frequency is:

```
[1] 0.04379186
```

This is shown in the descriptive analysis. Most of the undercounted votes lie between 0.00 and 0.10 percent range with 0.10 to 0.20 percent having the least number of undercounted votes.

## 6.4 Distribution of Percent Undercount

The distribution of the undercounted votes as a percentage for every county in Georgia is shown and its frequency relayed on a plot diagram




This is shown in the descriptive analysis.The frequency is normally distributed which is skewed to the right as it has a peak to the left of center. Most of the undercounted votes lie between 0.00 and 0.10 percent range with 0.10 to 0.20 percent having the least number of undercounted votes. The mean is at 0.04 percent as seen in the descriptive statistics. The rug function is shown at the bottom which creates a set of tick marks along the base of a plot to represent the concentration of the mean as a percentage. The rug is a one-dimensional display that is added to the existing plots to illuminate the mean of undercounted votesper county as expressed as a percentage. It represents values of the mean frequency by putting a symbol at various points along an axis.

## 6.5 Pairwise relationships

The relationship between pairs is shown in a table and a plot below.

## 6.6 Correlation Coefficients

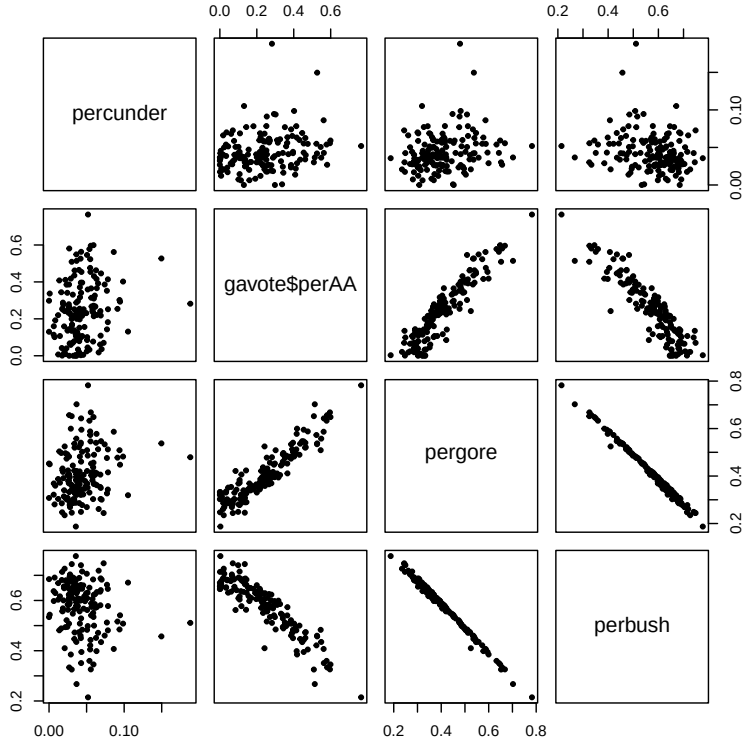
Table: Correlation Coefficients

| | perAA | gore | bush | other | votes | ballots |
|---|---|---|---|---|---|---|
| perAA | 1.0000 | 0.1314 | -0.0915 | -0.0278 | 0.0234 | 0.0277 |
| gore | 0.1314 | 1.0000 | 0.8086 | 0.8769 | 0.9531 | 0.9583 |
| bush | -0.0915 | 0.8086 | 1.0000 | 0.9493 | 0.9487 | 0.9428 |
| other | -0.0278 | 0.8769 | 0.9493 | 1.0000 | 0.9613 | 0.9561 |
| votes | 0.0234 | 0.9531 | 0.9487 | 0.9613 | 1.0000 | 0.9997 |
| ballots | 0.0277 | 0.9583 | 0.9428 | 0.9561 | 0.9997 | 1.0000 |

The correlation coefficient between perAA and itself is 1, which is expected as it is the correlation of a variable with itself. The positive correlation coefficient of 0.131 between perAA and gore suggests a weak positive relationship between the percentage of African Americans in a county and the number of votes for Gore in the 2000 US presidential election. On the other hand, the negative correlation coefficient of -0.091 between perAA and bush suggests a weak negative relationship between the percentage of African Americans in a county and the number of votes for Bush in the same election. The correlations between perAA and other variables, votes, and ballots are both positive but weak, suggesting a small positive relationship between the percentage of African Americans in a county and these variables.

The results from Table 6.6 can be displayed in a scatterplot as shown below. A pairwise scatter plot allows us to see the relationship between any two

variables of the concerned data-set. The table below shows the correlation relationship:




The diagonal shows the names of the four numeric variables of our data. The other cells of the plot matrix show a scatterplot (i.e. correlation plot) of each variable combination of our data frame

The table below shows the correlation coefficients between five different voting methods, namely LEVER, OS-CC, OS-PC, PAPER, and PUNCH, and the income levels of the voters. The income levels are classified into three categories: middle, poor, and rich.

Table: Economic conditions in comparison to Equipment used

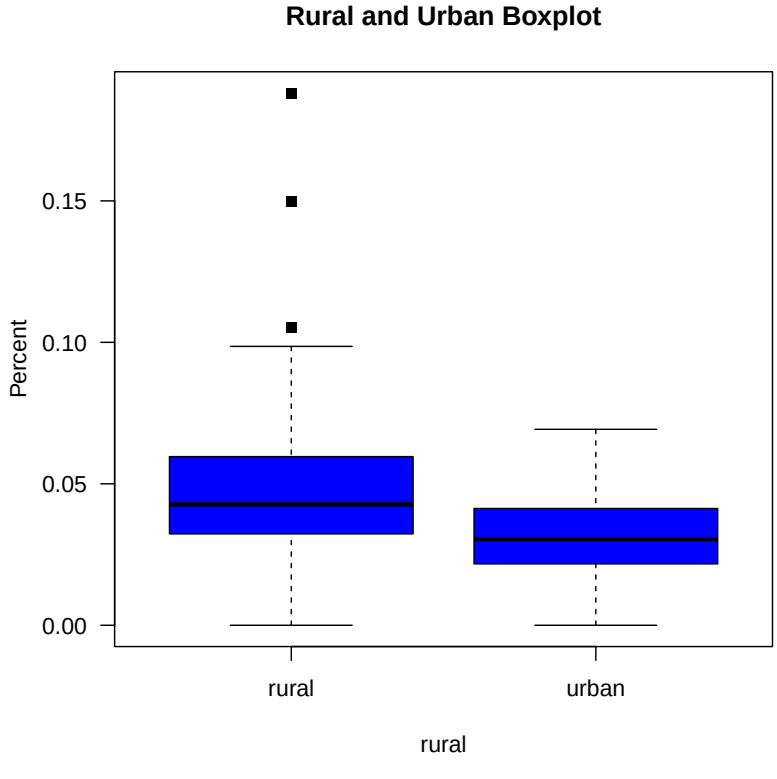
| | LEVER | OS-CC | OS-PC | PAPER | PUNCH |
|:---|:---:|:---:|:---:|:---:|:---:|
|middle | 0.035 | 0.049 | 0.043 | 0.006 | 0.046 |
|poor | 0.053 | 0.056 | 0.107 | 0.024 | 0.054 |

```
|rich    | 0.022 | 0.040 | 0.017 | -0.007 | 0.050 |
```

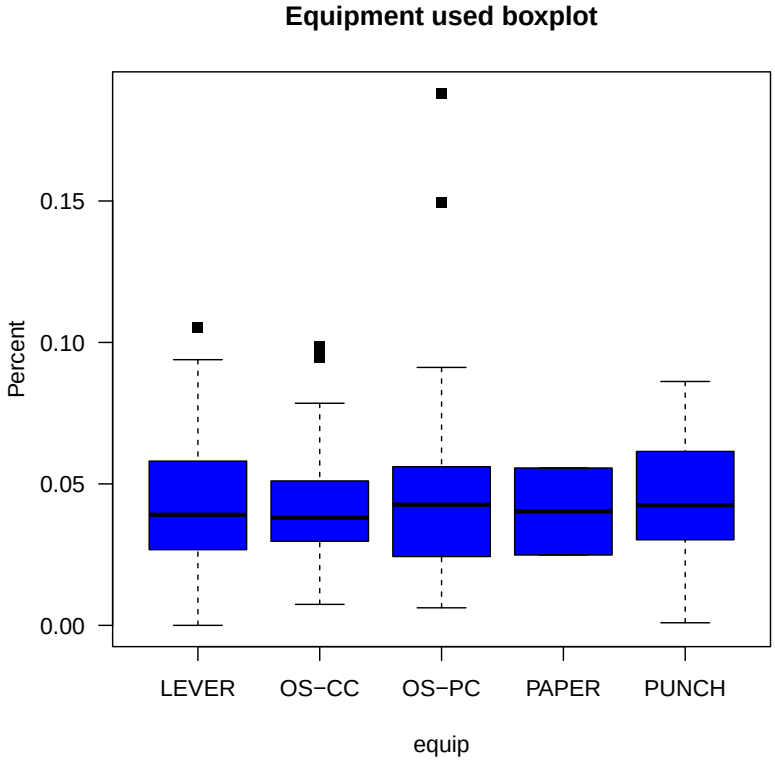
The correlation coefficients range from -0.007 to 0.107. The highest correlation coefficient of 0.107 is observed between the OS-PC voting method and the poor income level, indicating a stronger association between these two variables compared to other voting methods and income levels. In contrast, the PAPER voting method shows the weakest correlation with income levels, with correlation coefficients ranging from -0.007 to 0.024. Overall, the table suggests that income levels have a weak to moderate correlation with different voting methods.

## 6.7 Box plot

In descriptive statistics, box plot is a technique to show location, spread and skewness of numerical data groups through their quartiles.This relationship for both the rural and urban votes respectively is shown in the boxplot below:



The lines extending from the box show differences outside the upper and lower quartiles. The interquatile range is represented by the blue part, while the lines above and below the box show the skewness of the variable rural and urban respectively.




The box plot above represents that of the different technological equipments used during the voting process. This include lever, os-cc, os-pc, paper and punch. The interquatile range is represented by the blue part, while the lines above and below the box show the skewness of the each equipment.

# 7 Hypothesis Testing

- **Hypothesis**: There is a significant relationship between the undercount of votes and the explanatory variables of percentage of votes for Al Gore, percentage of African American population, usage of voting equipment, economic status, and total number of votes cast in a

county in Georgia.

- **Null Hypothesis**: There is no significant relationship between these variables and the undercount of votes.

This section presents the hypothesis and null hypothesis in list form. The hypothesis states that there is a significant relationship between the undercount of votes and the explanatory variables, which include the percentage of votes for Al Gore, percentage of African American population, usage of voting equipment, economic status, and total number of votes cast in a county in Georgia. The null hypothesis, on the other hand, suggests that there is no significant relationship between these variables and the undercount of votes.

# 8 Exploratory analysis

To understand the relationship between the dependent and independent variables in a linear model requires conducting exploratory analysis.Exploratory analysis is the relationship between the dependent and independent variables to show the regression analysis of the fitted linear model The response variable is undercount while the explanatory variables are pergore, perAA, usage, econ and gavote.

## Linear Regression Model

The lm function is used to fit linear models. lm returns an object of class "lm" to show important details about the connection between the response variable and the explanatory factors.

$$Y = \beta_{\mathbf{0}} + \beta_{\mathbf{1}}\mathbf{X_1} + \beta_{\mathbf{2}}\mathbf{X_2} + \beta_{\mathbf{3}}\mathbf{X_3} + \beta_{\mathbf{4}}\mathbf{X_4} + \beta_{\mathbf{5}}\mathbf{X_5} + \epsilon$$

Where:

Y: Response variable (Undercount)

X1: Number of votes for Gore

X2: Percent of African Americans in county

X3: Indicator of whether county is rural or urban

X4: Economic status of county poor

X5: Indicator variable of whether the county is rich

$\beta_{\mathbf{0}}$, $\beta_{\mathbf{1}}$, $\beta_{\mathbf{2}}$, $\beta_{\mathbf{3}}$, $\beta_{\mathbf{4}}$, $\beta_{\mathbf{5}}$ represent Coefficients of respective explanatory variables on response variable and $\epsilon$ represent the error term

To be more specific:

$\mathbf{undercount} = \beta_0 + \beta_1\mathbf{pergore} + \beta_2\mathbf{perAA} + \beta_3\mathbf{ruralurban} + \beta_4\mathbf{econpoor} + \beta_5\mathbf{econrich} + \epsilon$

Regression Results

| | Dependent variable: |
|---|---|
| | percunder |
| pergore | 0.016 (0.045) |
| perAA | -0.012 (0.030) |
| rural | -0.004 (0.005) |
| econ | 0.017*** (0.005) |
| econrich | -0.013* (0.007) |
| Constant | 0.035*** (0.013) |
| Observations | 159 |

```
R2                              0.200
Adjusted R2                     0.173
Residual Std. Error      0.023 (df = 153)
F Statistic           7.632*** (df = 5; 153)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
=========================================================
(Intercept) pergore perAA  ruralurban econpoor econrich
---------------------------------------------------------
0.035        0.016  -0.012   -0.004    0.017    -0.013
---------------------------------------------------------
```

The same model with the coefficients:

$$\textbf{undercount} = 0.035 + 0.016\textbf{pergore} - 0.012\textbf{perAA} - 0.004\textbf{ruralurban} + 0.017\textbf{econpoor} - 0.013\textbf{econri}$$

F-statistic is adopted to test the regression model. The p-value is

```
       value
1.987827e-06
```

and therefore, the null hypothesis is not accepted because the p-value is significantly different from zero and less than the chosen significance level of 99 percent. This means the alternative hypothesis is accepted.The explanatory variables have an impact on the undercount of votes

# 9 Conclusion

According to the report, undercounting of ballots was more prevalent in Georgia's poorest and rural areas. Depending on the tools utilized, the percentage

of African American voters had various implications on undercounting. Notably, with a rise in the number of voters who are African Americans, the lever method seems to be the most successful in decreasing undercounting. Although the adoption of voting computer technology was advocated following the election's technological failure, the study has revealed that paper ballots were more effective at the time. This was probably because voters were more accustomed to using paper votes than punch cards, which were less common and would have confused them. These results have significant implications for election officials and policymakers in the ongoing efforts to improve the accuracy and reliability of the voting process, particularly in historically disenfranchised communities.

**Reference List**

Meyer, M. (2002). *Uncounted Votes: Does Voting Equipment Matter? Chance methods.* 15(4), 33-38 Publications. Hlavac, Marek (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables.* R package version 5.2.3. https://CRAN.R-project.org/package=stargazer