Voting Data Analysis Using Frequentist and Bayesian Models

Ricky Li, Cody Seaman, Freddy V., Molly Jaroneski
STT465

Introduction

In this project we are analyzing the dataset “gavote” in the R-package “faraway”, which is a dataset containing information on the voters in Georgia during the 2000 Presidential election. This specific dataset was gathered due to the controversy surrounding whether or not counting machines were counting accurately. We wish to discover what variables have the most impact on undercount. We also want to know which model is best at predicting undercount in the Bayesian and Frequentist methods. After our analysis we hope to be able to accurately predict the undercount for the state of Georgia.

For each of the 159 counties in Georgia, data is taken after the voting happens. So, each observation contains the data for only that one county’s voting. The names and interpretations of the variables are as follows:

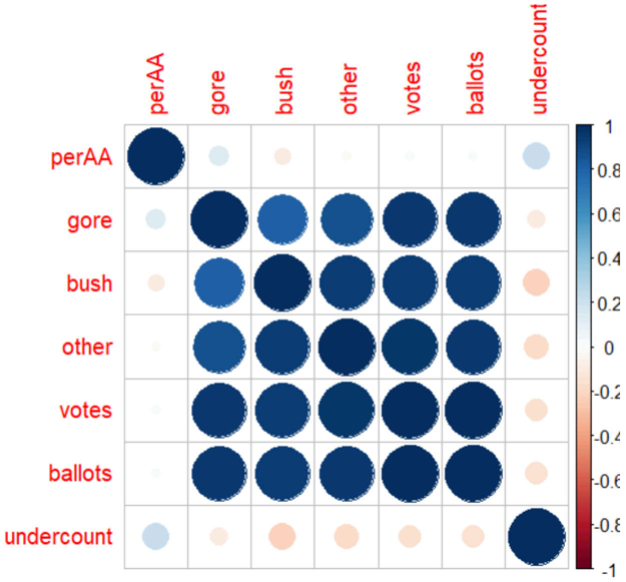
- Equip: The type of equipment used to conduct voting, this is a categorical variable. The types of vote counting used are level machine voting, optical scan central-count, optical scan precinct-count, paper ballots, and a punch card voting system.
- Econ: Economic status, also a categorical variable. The categories are middle class, poor, and rich.
- PerAA: Percentage of African American population. The values for this will be in percentages.
- Rural: A binary indicator of whether or not an area is rural. 0 is rural, 1 is urban.
- Atlanta: A binary indicator of whether or not an area is in Atlanta. 0 is in Atlanta, 1 is outside of Atlanta.
- Gore: Total votes for Al Gore for president in the form of a count variable.
- Bush: Total votes for George W. Bush for president in the form of a count variable.
- Other: Total votes for other candidates in the form of a count variable.
- Votes: Total number of votes in the form of a count variable.
- Ballots: Total number of ballots in the form of a count variable.

Exploratory Data Analysis

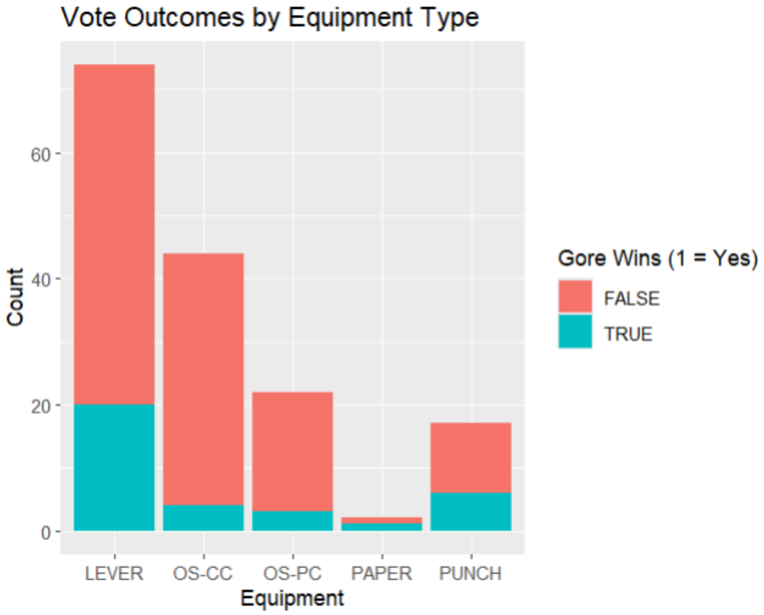
Taking a closer look at the dataset and the value types for the variables we see that we are only dealing with integer and factor types. Also, the dataset itself is not “extremely large” as we only have a total of 159 observations of 10 variables. This gives us only 1,590 data points.

Since the gavote dataset doesn’t contain any missing/NA values, we do not have to deal with missing data points. We do however, need to create the target variable undercount, which is calculated using the equation $undercount = (ballots - votes) / ballots$. This variable is a proportion which takes values in the range of $[0, 1]$. After the undercount variable column is added to the dataframe, all pre-processing and data cleaning is done.

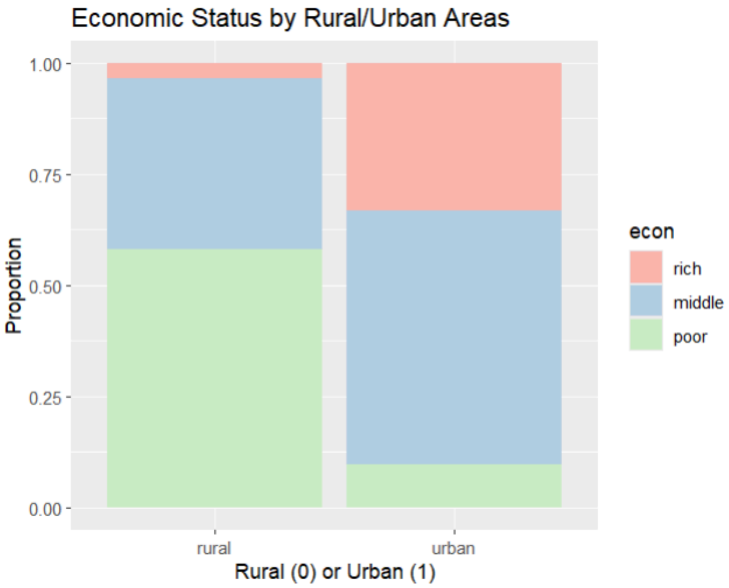
Now that we have the undercount value added to our dataset we can continue with some exploratory analysis.



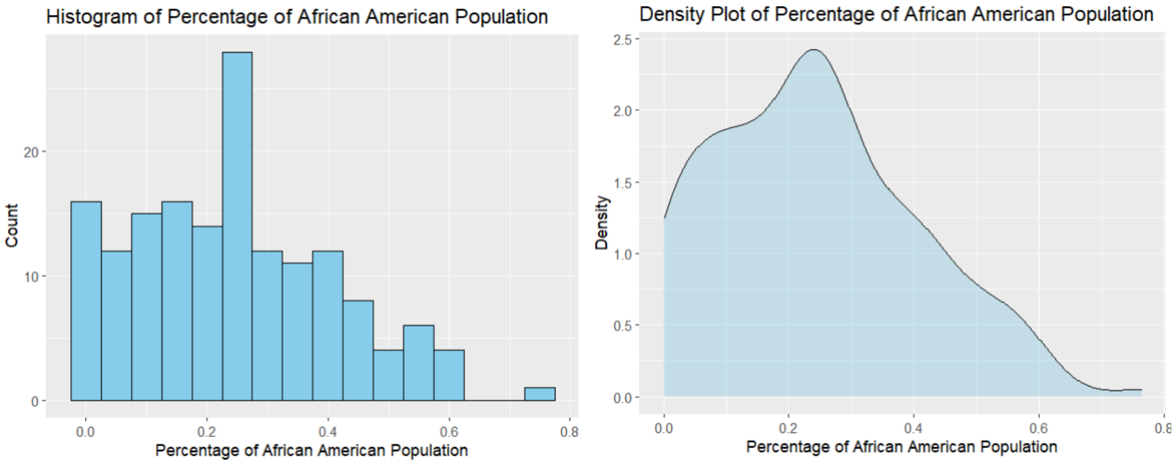
From first glance the correlation matrix isn't too helpful. Many of the highly correlated values here are of no surprise, if anything it is nice to see that undercount has between a -.4 & .5 correlation with all the other variables.



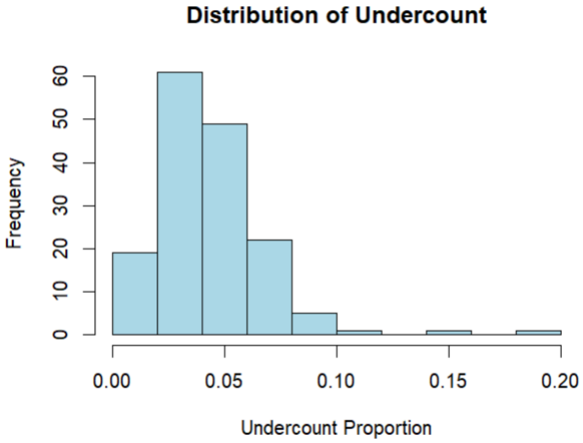
Shown above is a count of votes for each of the presidential candidates on each type of voting equipment. Its purpose is to visualize whether Gore or Bush would win by voting types. These help to recognize potential voting outcomes by equipment. As shown above, Al Gore winning was highly unlikely when looking at the graph. Bush seems to have a great majority of the votes across all types of equipment except paper. Paper voting is the only time when bush and gore have similar chances, but even then all other types of equipment are much more valuable. We can also observe that voting via lever was by far the most popular method.

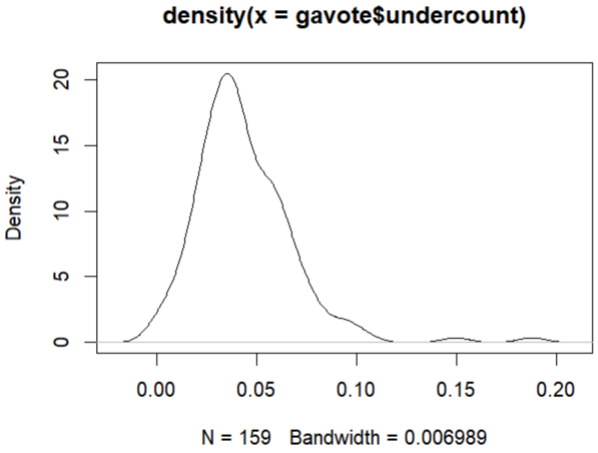


Another important factor to take into account when looking at voter demographics is of course the economic backgrounds of which voters are coming from for a certain area. Shown above is a split between those who fall into upper, middle, and lower class economic graphics along with where these people live. We can observe that in rural areas, more than half of its voters were part of the poor economic class, whereas in urban areas, the middle and rich economic class make up most of the voting population.



Next shown above are both histograms for percentage of African American Population along with respective distributive values in a density plot. The histogram helps to show how many counts there are for areas with their respective population percentage. And the density graph helps to visualize how likely a certain percentage for a population is given its spread. The density plot gets cut off at around .75, but it is reasonable to assume values above this would have a very low probability and not be 0. To reiterate, The graph on the left, histogram of percentage of African American population represents the frequency in which these percentages appear. The density plot of percentage of African American population on the right shows the given probability for a percentage of African American population.





Looking at the histogram and density plots for the undercount variable we can observe that the distribution follows a bell-shaped curve with a few outliers. This suggests that the variable undercount is approximately normally distributed. The density graph shows 2 cuts between 0.10 and 0.20 for which the values are assumed to be very low but not 0.

Methods

The first method we will implement is the frequentist regression model using all possible explanatory variables. Given that the response variable, undercount, is a continuous proportion, we can model the relationship between the explanatory variables and undercount using a linear regression model, lm(). Interaction terms are not included in any of the models, as adding interaction terms when variables are highly correlated (observed from correlation plot) results in increasing multicollinearity. Avoiding interaction terms also helps maintain interpretability.

Next, we will implement the reduced model using two methods, backward selection with the built-in step function and backward selection using the backselect.R file provided in Lecture 6. Both methods initially start with the full model, but the main difference is that the built-in step function iteratively removes variables one at a time until reaching the minimum AIC value. Whereas the backselect method iteratively removes the variable with the highest p-value if that p-value exceeds a threshold (default threshold is 0.05). The backselect method also expects the inputs to be standardized. The reduced models allow us to reduce and select only the most important variables that impact the prediction of undercount.

For the final model, we will fit a normal Bayesian model using Bayesian model selection and Bayesian model averaging. It is important to note that using the full set of variables for MCMC results in a nearly singular matrix due to linearly dependent variables which causes many problems, so to handle this, we will use only the selected variables obtained from the back select model to reduce the number of explanatory variables.

Figure 1: Regression model using all explanatory variables.

```
Call:
lm(formula = undercount ~ ., data = gavote)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.060792 -0.012180 -0.001271  0.009021  0.103324 

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        2.976e-02  1.093e-02   2.723 0.007260 ** 
equipOS-CC         8.884e-03  4.231e-03   2.100 0.037482 *  
equipOS-PC         2.503e-02  5.508e-03   4.543 1.16e-05 ***
equipPAPER        -1.040e-02  1.492e-02  -0.697 0.486853    
equipPUNCH         1.237e-02  6.363e-03   1.945 0.053770 .  
econpoor           1.819e-02  4.525e-03   4.019 9.34e-05 ***
econrich          -1.083e-02  7.919e-03  -1.367 0.173616    
perAA              2.651e-03  1.382e-02   0.192 0.848128    
ruralurban        -4.585e-03  4.878e-03  -0.940 0.348801    
atlantanotAtlanta  2.130e-03  9.464e-03   0.225 0.822262    
gore              -1.253e-05  3.148e-06  -3.981 0.000108 ***
bush              -1.209e-05  2.938e-06  -4.115 6.48e-05 ***
other              2.754e-06  6.161e-06   0.447 0.655483    
votes                     NA         NA      NA       NA    
ballots            1.143e-05  2.806e-06   4.072 7.64e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02067 on 145 degrees of freedom
Multiple R-squared:  0.3711,	Adjusted R-squared:  0.3147 
F-statistic:  6.58 on 13 and 145 DF,  p-value: 9.236e-10

[1] "AIC:  -767.01429601246"
```

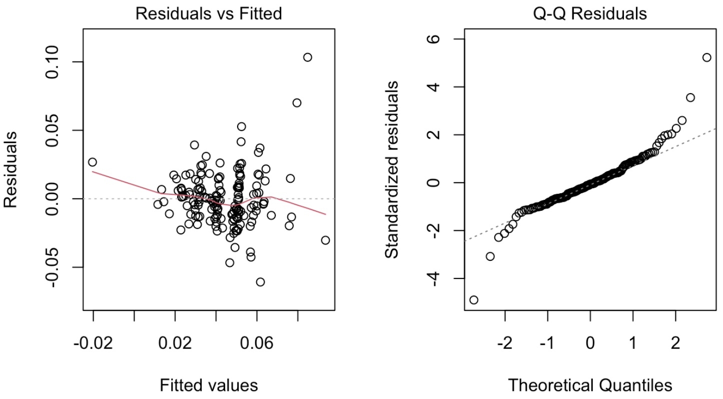


Figure 2: Reduced model using built-in backward step function.

```
Call:
lm(formula = undercount ~ equip + econ + gore + bush + ballots, 
    data = gavote)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.058880 -0.011935 -0.001180  0.008808  0.103648 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.092e-02  3.363e-03   9.192 3.13e-16 ***
equipOS-CC   8.401e-03  4.095e-03   2.052   0.0420 *  
equipOS-PC   2.472e-02  5.378e-03   4.598 9.05e-06 ***
equipPAPER  -1.009e-02  1.474e-02  -0.685   0.4947    
equipPUNCH   1.049e-02  6.039e-03   1.737   0.0845 .  
econpoor     1.989e-02  3.689e-03   5.392 2.67e-07 ***
econrich    -1.300e-02  6.432e-03  -2.020   0.0451 *  
gore        -1.223e-05  2.952e-06  -4.141 5.76e-05 ***
bush        -1.176e-05  2.722e-06  -4.322 2.82e-05 ***
ballots      1.119e-05  2.662e-06   4.205 4.48e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02048 on 149 degrees of freedom
Multiple R-squared:  0.3653,	Adjusted R-squared:  0.327 
F-statistic:  9.53 on 9 and 149 DF,  p-value: 2.123e-11

[1] "AIC:  -773.57291009901"
```

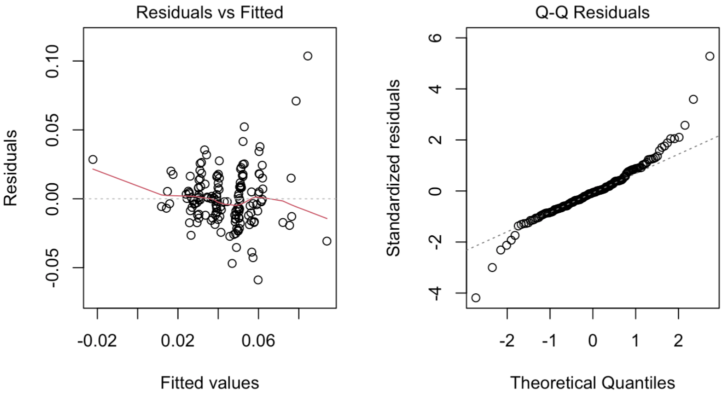
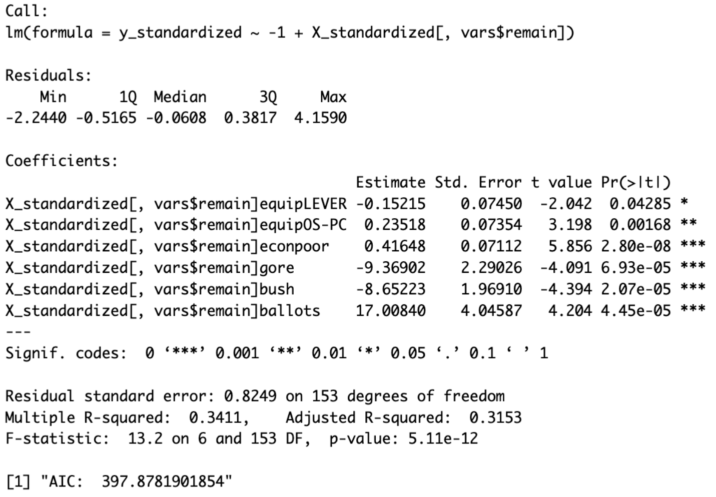


Figure 3: Reduced model using backward selection function in backselect.R (Lecture 6)



```
Call:
lm(formula = y_standardized ~ -1 + X_standardized[, vars$remain])

Residuals:
    Min      1Q  Median      3Q     Max 
-2.2440 -0.5165 -0.0608  0.3817  4.1590 

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
X_standardized[, vars$remain]equipLEVER -0.15215    0.07450  -2.042  0.04285 *  
X_standardized[, vars$remain]equipOS-PC  0.23518    0.07354   3.198  0.00168 ** 
X_standardized[, vars$remain]econpoor    0.41648    0.07112   5.856 2.80e-08 ***
X_standardized[, vars$remain]gore       -9.36902    2.29026  -4.091 6.93e-05 ***
X_standardized[, vars$remain]bush       -8.65223    1.96910  -4.394 2.07e-05 ***
X_standardized[, vars$remain]ballots    17.00840    4.04587   4.204 4.45e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8249 on 153 degrees of freedom
Multiple R-squared:  0.3411,	Adjusted R-squared:  0.3153 
F-statistic:  13.2 on 6 and 153 DF,  p-value: 5.11e-12

[1] "AIC:  397.8781901854"
```

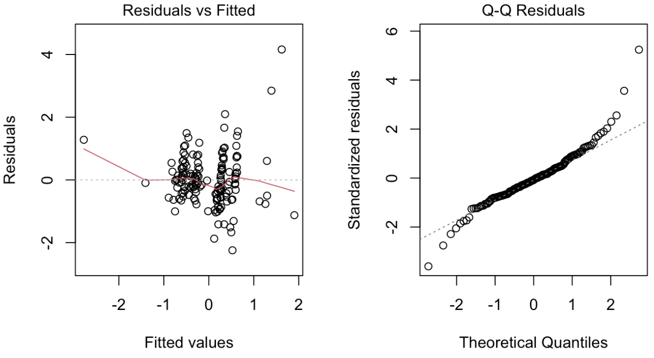
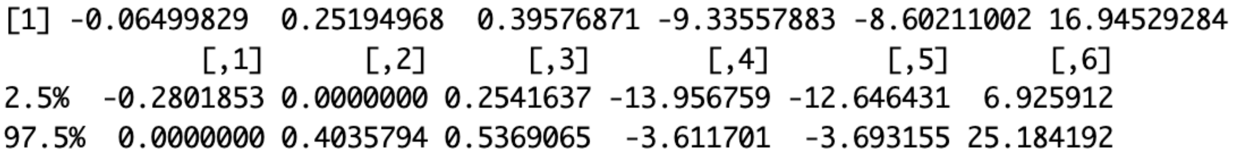


Figure 4: Posterior coefficient means and 95% credible intervals of selected variables (from back select model) using Gibbs Sampler



```
[1] -0.06499829  0.25194968  0.39576871 -9.33557883 -8.60211002 16.94529284
             [,1]      [,2]      [,3]       [,4]       [,5]      [,6]
2.5%   -0.2801853 0.0000000 0.2541637 -13.956759 -12.646431  6.925912
97.5%   0.0000000 0.4035794 0.5369065  -3.611701  -3.693155 25.184192
```

Figure 5: Trace plots for significant beta coefficients

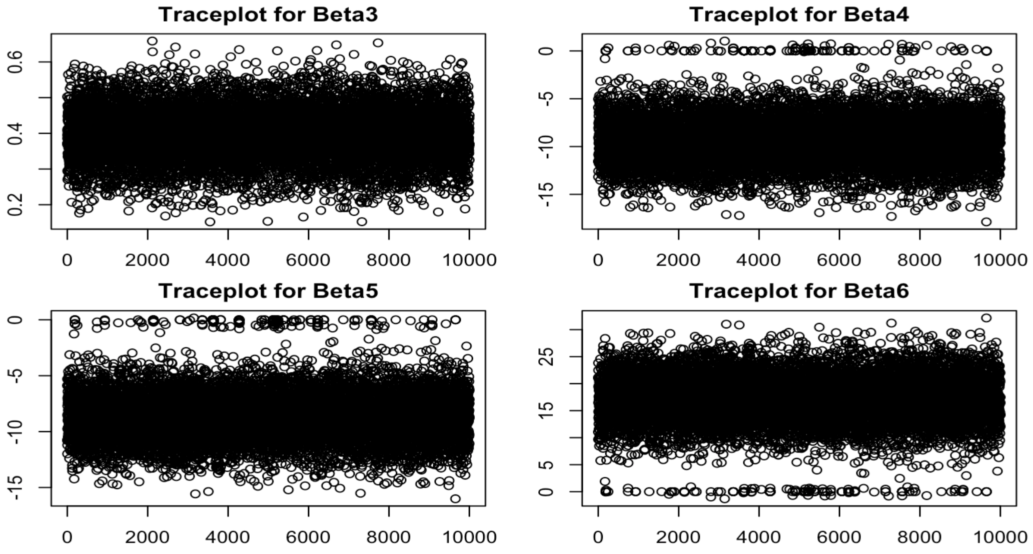


Figure 6: ACF plots for significant beta coefficients

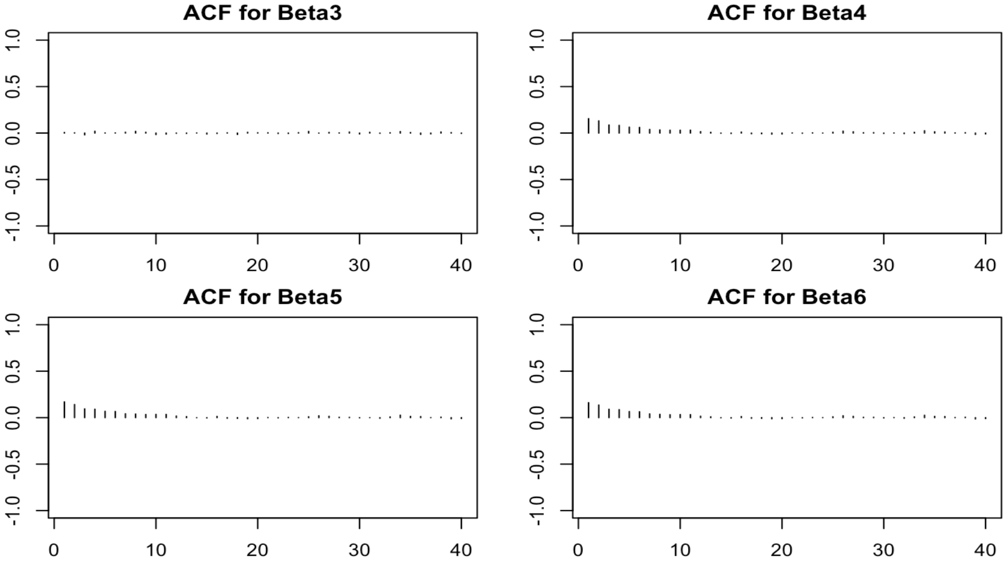
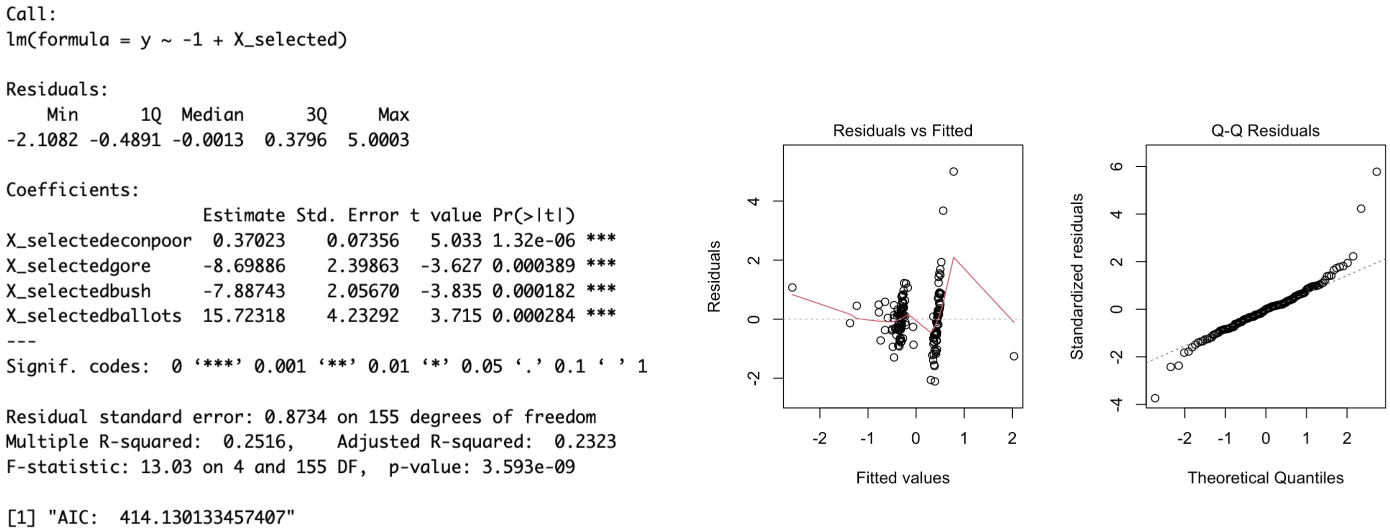


Figure 7: Bayesian Model obtained from Bayesian Model Selection/Bayesian Model Averaging



```
Call:
lm(formula = y ~ -1 + X_selected)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1082 -0.4891 -0.0013  0.3796  5.0003 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
X_selectedeconpoor  0.37023    0.07356   5.033 1.32e-06 ***
X_selectedgore     -8.69886    2.39863  -3.627 0.000389 ***
X_selectedbush     -7.88743    2.05670  -3.835 0.000182 ***
X_selectedballots  15.72318    4.23292   3.715 0.000284 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8734 on 155 degrees of freedom
Multiple R-squared:  0.2516,	Adjusted R-squared:  0.2323 
F-statistic: 13.03 on 4 and 155 DF,  p-value: 3.593e-09

[1] "AIC:  414.130133457407"
```

Model Diagnosis and Variable Selection

We can see that in the full model (Figure 1), the variables that impact undercount the most are the equipment used to vote, the economic status of the county, who the citizens voted for, and the number of ballots. The variables that don't impact undercount the most would be perAA (the percentage of African Americans), and whether or not the county is within Atlanta. In the residual plot we can see that the residuals are somewhat uniformly scattered around the zero line with no obvious patterns and only a couple outliers. This suggests that the relationship between the response variable and predictors is mainly linear. Our Q-Q plot has some weird angling towards both ends which reflect the outliers, however, the residuals generally follow the diagonal reference line, suggesting that the residuals are approximately normal. Since the AIC value associated with the model is -767.01, we can conclude that the full model is fairly balanced, but can be improved since we observe predictors with minimal/no impact on undercount.

Looking at Figure 2 and Figure 3, we can see that both models show us what we expected to be the most impactful variables for the prediction of undercount. However, the second model, using the backselect.R file, reduces it further to the exact factors inside of variables that impact prediction the most. One such example is the econpoor factor of the econ variable, suggesting that being a poor community affects undercount more than a rich one. This can also be seen with the type of equipment used to vote, such as equipLEVER (level machine voting), equipOS-PC (optical scan precinct-count). The residual plots for both models look to be more condensed than the full model, however the second model looks to have less spread. The Q-Q plot for the first reduced model looks similar to the full model, however the second reduced plot looks more linear. Looking at the AIC values for both reduced models, we can observe that the model using the built-in step function results in an AIC value of -773.57 whereas the model using the back select function results in an AIC value of 397.88. When comparing models using AIC, the model with a lower AIC value indicates a better fit so we would choose the built-in step function model as the better model.

For the model in the Bayesian framework, we perform Bayesian variable selection and model averaging using a g-prior. The response variable undercount follows a normal distribution and its likelihood is given by $y \sim N(X^T\beta,\ \sigma^2 I)$. The inclusion indicators for each variable follow a Bernoulli prior with equal probability given by $z_j \sim Bernoulli(0.5)$. We use this prior to reflect the non-informative prior knowledge. For each iteration of MCMC, we sample and update z_j, then sample the regression coefficients β where the variable is included $(z_j = 1)$ from the posterior distribution under the g-prior. As noted above, we start with a reduced X matrix to resolve the computational problem of a singular matrix. This reduced X matrix uses the variable selected by the backward selection model which are equipLEVER, equipOS-PC, econpoor, gore, bush, ballots, and β_1, ..., β_6 correspond to each variable's coefficient respectively.

The output in Figure 4 are the results of the Bayesian model selection and model averaging. We can see in observe that β_1 and β_2 contain zero in their 95% credible intervals which suggest that these variables are not significant in our model, and therefore will not be included. Looking at the trace plots in Figure 5, we can observe that the significant coefficients (β_3, β_4, β_5, β_6) quickly converge and reach equilibrium and their ACF plots in Figure 6 reduce to zero as the lag increases which suggest little to no significant correlation between the variables. We can see from the Bayesian model summary in Figure 7 that all of our variables are significant and that the Q-Q plot follows the diagonal reference line very closely except towards both ends. The residual plot also has residuals somewhat uniformly scattered around zero, however there seems to be some clustering happening which suggests that the Bayesian model isn't capturing an underlying trend in the data. The model also produces an AIC score of 414.13 which is the worst AIC score out of all the models we have created. However, we should take into account that the Bayesian model offers an advantage in variable selection as the model only includes variables whose credible intervals exclude zero leading to a smaller and more interpretable subset of variables. This subset is valuable for understanding the most important relationships in predicting undercount in the dataset.

Conclusion

After implementing several different regression models in both Frequentist and Bayesian frameworks, it can be concluded that the reduced model obtained from the built-in step function is the best at predicting undercount. This is mostly because this model achieved an AIC value of -773.57 which is the lowest of all the models. This model's residual plot has its residuals uniformly distributed around the zero line with no obvious underlying trend, suggesting that the model is able to fit the target variable undercount fairly well. The model's Q-Q plot has its residuals mainly following the diagonal reference line which indicate that the residuals are normally distributed.

In terms of model complexity and interpretability, the Bayesian model obtained from Bayesian model selection and averaging is the best as the model contains the least number of explanatory variables while still maintaining a decent model AIC score relative to the other reduced models. The Bayesian model incorporates non-informative prior information on the inclusion indicators and is able to produce the posterior coefficient information for each variable such as posterior means and credible intervals allowing for more informative inference compared to the Frequentist models. The use of the posterior coefficient information also reduces the risk of overfitting and provides a more robust method of variable selection.

The main difference between the models implemented in a Frequentist framework and a Bayesian framework deals with their approach to variable inference. For Frequentist models, we focus on reducing the AIC score while maintaining/increasing predictive power based solely on the observed data. This often results in models with more explanatory variables, larger complexity, and lower interpretability. For Bayesian models, we incorporate prior information (non-informative priors in our case) to obtain the posterior distributions of coefficients which allow us to use inferential statistics such as credible intervals and posterior means which are more interpretable than confidence intervals and point estimates. This often allows us to generate models with less explanatory variables while still maintaining predictive performance.