STT 465 Final Report:

Examining a PRIOR Election: A Look into 2000 Elections Using OLS and Bayesian Techniques

Sam Isken

December 9, 2019

Note

This PDF contains the written report (no code visible). The other RMarkDown included contains the code as well as parts of the essay portions of this project (ones that are relevant to the coding scheme). I have also included a supplemental dataset I used to examine the election results at the county level of granularity. This is over the 8 page limit only due to visualizations and LaTeX Code.

Introduction

The 2000 United States Presidential Election was bound to be contentious. George W. Bush, a conservative with a family legacy in the political sphere was running against incumbent Vice President (to Moderate Democrat Bill Clinton) Al Gore. Both ran campaigns focused on domestic issues with some topics of foreign policy sprinkled in. The conservatives emphasized "family values" after Clinton's scandal regarding Monica Lewinsky while Gore criticized Bush's lack of political experience and conservative policy regarding Healthcare.

After election night it was clear this election would not go as smoothly as planned. The biggest story to come out was that of the "hanging chad" which made voters in Florida unsure of for whom they were casting their ballot and lead to the famous Florida recount (which was interrupted when a conservative group stormed the recount offices, led by recently convicted felon Roger Stone). This led to a general skepticism on the part of the American people about the security and fairness of the election results. This was eventually decided in the SCOTUS, which handed the election to Bush.

This paper will focus on another election night controversy, the state of Georgia. We are looking to examine the undercount of votes granularly at the county level. The undercount is defined as the difference between the number of ballots cast and votes recorded for president. Voters may have chosen not to vote for president, voted for more than one candidate (disqualified) or the equipment may have failed to register their choice. (**Source**: "*{r} help(gavote)*"). Our goal is to determine the factors that significantly predict the undercount at the county level, which in turn could be used to make inferences for future elections. Our data contains the following variables to assist with this prediction:

Data Summary

- **equip**: The voting equipment used: LEVER, OS-CC (optical, central count), OS-PC (optical, precinct count) PAPER, 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
- **undercount**: (number of ballots - number of votes) / (number of ballots)

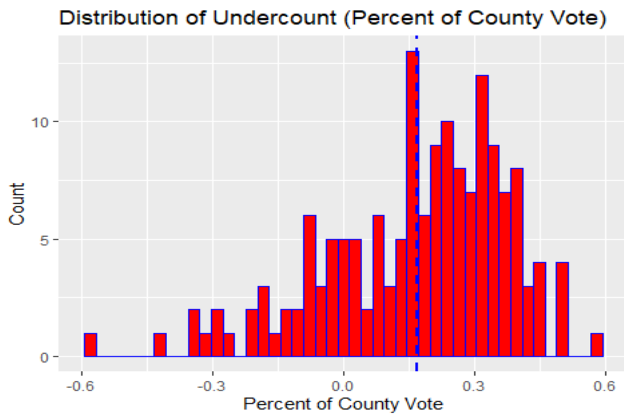
(**Source**: "*Meyer M. (2002) Uncounted Votes: Does Voting Equipment Matter? Chance, 15(4), 33-38 & "{r} help(gavote)")*

Surface-Level Data Analysis

From a surface level analysis, it was clear that in some counties, up to 18% of ballots were thrown out in some counties due to voting or machine errors. (This was calculated using various functions in R). The average percent of votes thrown out by county was 4%. This quick look proved that this is an important topic as this easily could have swung counties in favor of the other candidate (either Gore or Bush). This provided justification for our research.

A new data set was then loaded, containing election results at the county level for Georgia. This assisted in providing a basis / set of background knowledge for our project.

The visual below plots the distribution of undercount. The percent of county vote is on the x-axis and the number of counties for which this value is true is on the x-axis. From this it can be shown that the average level of undercount was 4% and that for some counties large percentages of votes were removed due to perceived errors.



It can be seen that 45% of county results were within the margin of the undercount (18%). This, coupled with the fact that Bush only won by a relatively narrow margin (54.67% to 42.98%) statewide, prompted the analysis of this data.

(**Source**: *https://en.wikipedia.org/wiki/2000_United_States_presidential_election_in_Georgia*)

This concluded the argument in favor (justification) of conducting this study. The paper will next present the final models (both OLS and Bayesian) and methods used to create them.

Final Models

OLS Frequentist Regression

$$\text{UNDERCOUNT} = 5.807496 * 10^{-3} x_1 + 2.273752 * 10^{-2} x_2 - 9.044148 * 10^{-3} x_3 + 6.082960 * 10^{-3} x_4 + 2.089696 * 10^{-2} x_5 - 1.894076 * 10^{-2} x_6 - 3.554025 * 10^{-5} x_7 - 3.386005 * 10^{-5} x_8 + 3.282120 * 10^{-5} x_9 + 2.898298 * 10^{-2} + \epsilon_i$$

$$\text{UNDERCOUNT} = \beta_1 + \text{EQUIP}x_1 + \text{ECON}x_2 + \text{GORE}x_3 + \text{BUSH}x_4 + \text{BALLOTS}x_5 + \epsilon_i$$

The above functions show both the specific numerical coefficients within the regression function as well as the generic variables they represent. Because both EQUIP and ECON are indicator variables, indicating a specific voting machine or economic status of a county, multiple coefficients are created for each.

Bayesian Analysis (Normal Model)

Within a Bayesian Analysis (applied using a Normal Model) the result is a prior, posteriors and a likelihood function:

Likelihood function:

$$\text{Likelihood Function of Undercount (Y):}$$
$$p\left(\text{UNDERCOUNT (Y)}|\text{EQUIP,ECON,GORE,BUSH,BALLOTS}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(\text{UNDERCOUNT}_i - \beta_1 - \text{EQUIP}_2 x_i - \text{ECON}_3 x_i - \text{GORE}_4 x_i - \text{BUSH}_5 x_i - \text{BALLOTS}_6 x_i)^2\right\}$$

Prior:

$$\text{Flat Prior:}$$
$$p(\beta_1, \ldots, \beta_p) \propto 1$$

Posterior Distributions:

Posterior Distribution of Regression Coefficients:

$$p\,(\mathrm{EQUIP,ECON,GORE,BUSH,BALLOTS}|\mathrm{UNDERCOUNT\ (Y)}) \propto 1 \times \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(\mathrm{UNDERCOUNT}_i - \beta_1 - \mathrm{EQUIP}_2 x_i - \mathrm{ECON}_3 x_i - \mathrm{GORE}_4 x_i - \mathrm{BUSH}_5 x_i - \mathrm{BALLOTS}_6 x_i)^2 \right\}$$

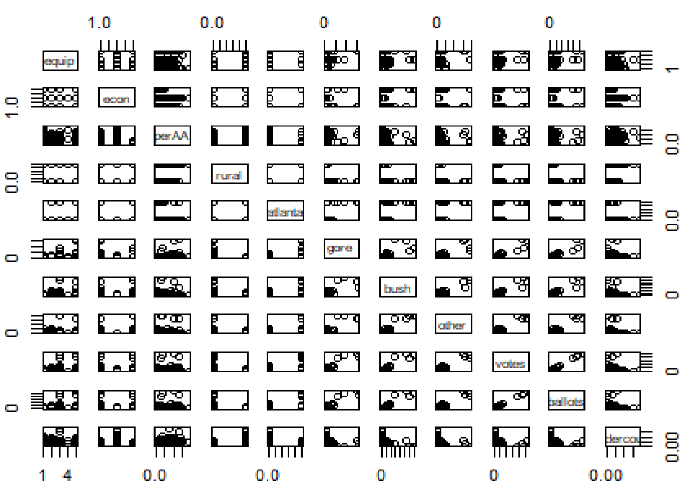
These results, and how they were derived (using both numeric methods as well as mathematical statistics) will be further explained in both the Methods and Conclusion sections.

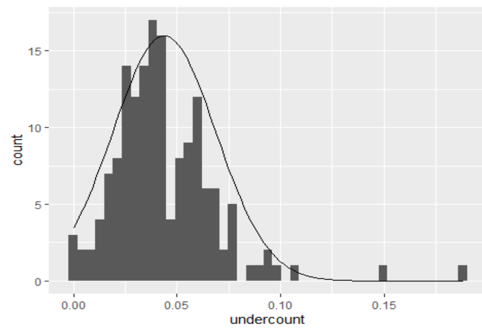
Methods

Two methods were applied to model this data. First, a basic Frequentist OLS regression model (lm() in R) along with forward-stepwise variable selections was used. This was used to perform regression diagnostics and make inferences on characteristics that could make counties prone to undercount.

Next, a Normal Bayesian model was fitted to the data. This model provided a likelihood function, priors and posterior distributions. Since there was no prior information (either expert or guess) on the priors a flat (non-informative prior) was applied. Upon reading about the application of priors and Bayesian Statistics in political science it was clear that creating priors for elections is incredibly difficult, especially priors regarding the ability of voting machines to function properly. Any priors applied within literature were typically normally distributed, educated guesses on the behavior of the data.

Exploratory Data Analysis





A deeper dive was then performed on the data. Two key observations from this (admittedly crowded) correlation matrix were the following:

1. Undercount is negatively correlated with the level of equipment.
2. Ballots and Votes are essentially co-linear.

This provided an insight into a variable to remove (either ballots or votes) as well as a variable that may predict undercount. The dependent variable was examined next in order to see how undercount was distributed vs. a normal distribution.

In order to make inference and run nessecary analysis (confidence intervals etc.) prior to Bayesian analysis it is desired for the dependent variable to be normally distributed. This was judged to be close enough (based on the histogram above). The dependent variable is roughly normal with a few right-side outliers.

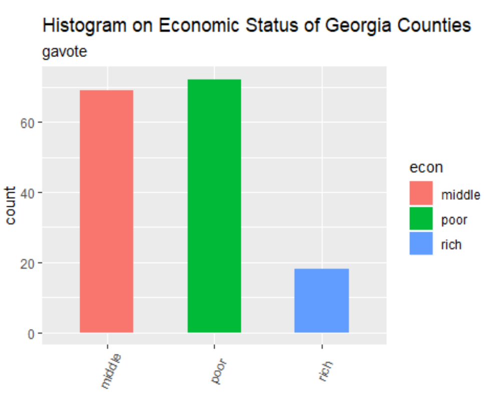
Shapiro-Wilk normality test

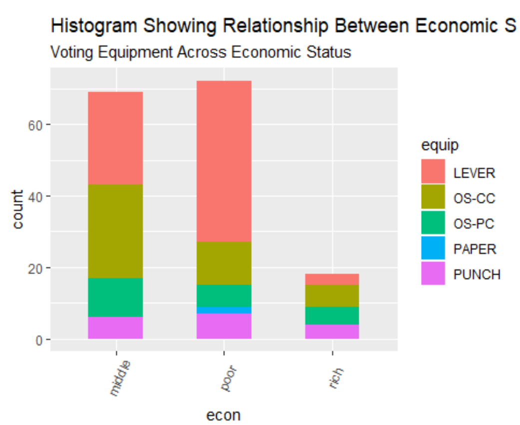
data: gavote$undercount

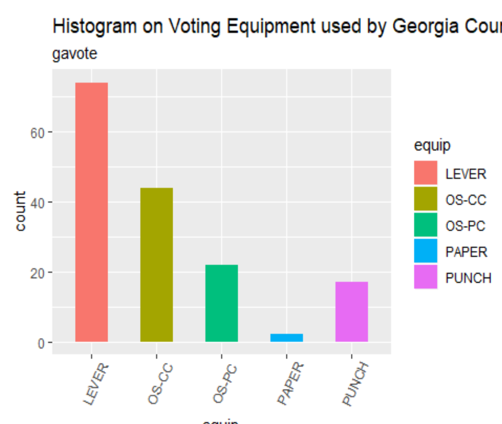
W = 0.88174, p-value = 6.176e-10

Though the sample did not pass the Shapiro-Wilk test this was acceptable as it is clear from the histogram that it is roughly normally distributed with some right-side outliers. The Shapiro-Wilk test often fails as N becomes large (in contrast to many other statistical tests). Normally, some of the outliers could be removed, however this was decided against due to the small dimensionality of the data.

Next, some plots regarding categorical variables were created to examine the possible relationship between economic status and the voting machines present in a county.







From these plots it was clear most counties in Georgia are poor and do not use the most reliable form of voting (paper). I believed it would be interesting to examine whether either of these categorical variables end up being significant in a model. A hypothesis was, that they would be (due to prior knowledge of politics and the history of voting) but they could be statistically insignificant in predicting undercount. It was also clear that there is a large increase in the use if lever machines for voting as economic status decreases within the data. Thus, if lever is found to be an unreliable form of voting, this could be used to make inference about which counties could have the least accurate voting counts.

The next section will discuss the methods for building both the OLS Frequentist Regression Model and the Bayesian Normal Model.

Model Building

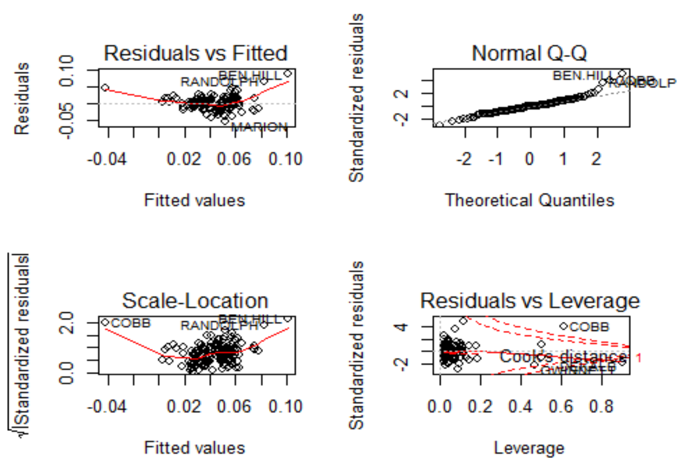
Frequentist OLS Regression Model, Variable Selection and Diagnostics

The frequentist model was created by first regressing undercount on every predictor. The output is shown below:

```
Call:
lm(formula = undercount ~ equip + econ + perAA + rural + atlanta +
    gore + bush + other + votes + ballots, 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
```

As one can see not nearly all the predictors are significant. Next, insignificant variables were removed (Performing Stepwise Variable Reduction) resulting in the following model:

```
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
```



Regression diagnostics were then performed. First a standard Rule of Thumb regarding outliers was applied: Outliers that have a very high leverage (greater than Cooks Distance) should be removed. This was done in order to preserve the size of the data.

The same model (with a single high-leverage outlier removed) was then re-run.

Next, regression diagnostics were performed:

1. **Residuals vs. Fitted**: The relationship (even after removing FULTON) is not very linear. This is shown by the curved line and lowers the confidence in this relationship being of the linear classification

2. **Normal Q-Q**: The residuals are roughly normally distributed with a few outliers. This is a positive sign in linearity for our model

3. **Scale-Location**: The data is very heteroskedastic (difference in variance across the residuals). This means the variance is very inconsistent across the data

4. **Residuals vs. Leverage**: Few points (after FULTON removal) lie outside of Cook's Distance

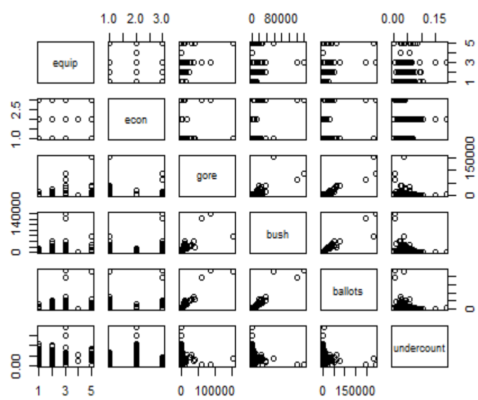
Overall, the predictors in this model were statistically significant. However, many of the regression diagnostics did not pass so the linearity of this relationship was in question. This was expected, as many situations within the social sciences (in which politics is included) are not very linear in their relationship.

Thus, our regression equation result is as follows:

$$\text{UNDERCOUNT} = 5.807496 * 10^{-3}x_1 + 2.273752 * 10^{-2}x_2 - 9.044148 * 10^{-3}x_3 + 6.082960 * 10^{-3}x_4 + 2.089696 * 10^{-2}x_5 - 1.894076 * 10^{-2}x_6 - 3.554025 * 10^{-5}x_7 - 3.386005 * 10^{-5}x_8 + 3.282120 * 10^{-5}x_9 + 2.898298 * 10^{-2} + \epsilon_i$$

$$\text{UNDERCOUNT} = \beta_1 + \text{EQUIP}x_1 + \text{ECON}x_2 + \text{GORE}x_3 + \text{BUSH}x_4 + \text{BALLOTS}x_5 + \epsilon_i$$

Next, the correlations between significant variables were examined. Undercount was found to be correlated with the class 'poor' in econ, but, suprisingly, not clearly correlated with any single type of election equipment.



Prediction probabilities to model likely scenarios were then created.

(This code chunk is shown so the reader can cleary see how these predictions were developed).

```
# Prediction Equation for: a poor county, with 3000 votes for Gore, 2000 votes for Bush, lever voting (0 for all) and 5500 ballots

y1=0+2.898298e-02+2.089696e-02+(3000*-3.554025e-05) +(2000*-3.386005e-05)+(5500*3.282120e-05)
print(paste("There is a predicted value",y1*100,"% of votes to be undercounted by in this county"))

## [1] "There is a predicted value 5.605569 % of votes to be undercounted by in this county"

# Prediction Equation for: a poor county, with 2000 votes for Gore, 3000 votes for Bush, lever voting (0 for all) and 5500 ballots
```

```
y2=0+2.898298e-02+2.089696e-02+(2000*-3.554025e-05) +(3000*-3.386005e-05)+(5500*3.282120e-05)
print(paste("There is a predicted value",y2*100,"% of votes to be undercounted by in this county"))

## [1] "There is a predicted value 5.773589 % of votes to be undercounted by in this county"

# Prediction Equation for: a RICH county, with 2000 votes for Gore, 3000 votes for Bush, lever voting (0 for all) and 5500 ballots


y3=0+2.898298e-02+-1.894076e-02+(3000*-3.554025e-05) +(2000*-3.386005e-05)+(5500*3.282120e-05)
print(paste("There is a predicted value",y3*100,"% of votes to be undercounted by in this county"))

## [1] "There is a predicted value 1.621797 % of votes to be undercounted by in this county"
```

The predicted percent of votes to be undercounted drops significantly in rich vs. poor economic status counties when other variables are held fixed. There does not appear to be a drastic change in undercount when the winning candidate is changed.

Overall, the linear model worked relatively well. It was able to isolate a decent number of significant predictors and allows a user to run various inputs to predict the demographics of counties that will experience undercount in other elections.

Normal Bayesian Model

Before running scripts to produce the Bayesian Normal Model the definition of what a Bayesian MLR is truly doing (behind the scenes of this function) was researched. The goal was to predict the undercount at the county level. From a Bayesian perspective, the goal was to get a posterior distribution of the possible model parameters based on the data and prior knowledge (from either a field expert or educated guess).

The MLR Model can be written as:

$$y_i = \sum_{j=1}^{p} x_{ij}\,\beta_j + \varepsilon_i$$

for

$$i = 1, \dots, \text{n}$$

This can be expressed in matrix form as:

$$\begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \cdot \\ \cdot \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix}$$

which can also be written as:

$$\mathbf{y} = \boldsymbol{\beta}^T \mathbf{X} + \boldsymbol{\varepsilon}$$

where

$$\beta\text{^}\{T\} \quad = \left(\beta_1 \cdots \beta_p\right)$$

is the transpose of the vector.

$$\boldsymbol{\beta} = \left(\beta_1 \cdots \beta_p\right)^T$$

The goal of OLS is to minimize:

$$S(\beta) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

From all of this we can obtain the following estimate that:

$$\hat{\beta}_{OLS} = (X^T X)^{-1}(X^T y)$$

All of this, comes with the normal assumption that

$$\varepsilon_1, \dots, \varepsilon_n \sim N(0, \sigma^2)$$

This can also be expressed in matrix form:

$$\varepsilon \sim \text{Multivariate}N(\mathbf{0}_{n\times\mathbf{1}}, \sigma_\epsilon^2 \mathbf{I}_{n\times n})$$

which implies:

$$\{y|X, \beta, \sigma^2\} \sim \text{ Multivariate } N(X\beta, \sigma^2 I)$$

Moving into the Bayesian perspective, the posterior distribution (beliefs after viewing data) can be defined as:

$$p(\theta|y) \propto p(\theta) * p(y|\theta)$$

(y = response data)

Within the model, since there is no expert opinion on priors, a uniform/flat prior were assumed. This means:

$$p(\beta_1, \dots, \beta_p) \propto 1$$

It is assumed the data comes from a normal population thus:

$$\epsilon_i \sim N(0, \sigma^2)$$

and thus:

$$y_i \; N(\beta_1 + \beta_2 x_i + \dots + \beta_n x_i)$$

From this it can be shown that the likelihood function is:

$$p(y_1, \dots, y_n|\beta_1, \dots, \beta_p) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \beta_1 - \beta_2 x_i - \dots - \beta_p x_i\right)^2\right\}$$

This can be derived and written in matrix form as:

$$\begin{aligned} p(y_1, y_2, \dots, y_n|\boldsymbol{\beta}, \sigma_\varepsilon^2) &= p(y_1|\boldsymbol{\beta}, \sigma_\varepsilon^2) \times \cdots \times p(y_n|\boldsymbol{\beta}, \sigma_\varepsilon^2) \\ &= \prod_{i=1}^{n} (2\pi\sigma_\varepsilon^2)^{-\frac{1}{2}} \mathrm{e}^{-\frac{\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)}{2\sigma_\varepsilon^2}} \\ &= (2\pi\sigma_\varepsilon^2)^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2} \\ &= (2\pi\sigma_\varepsilon^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})} \end{aligned}$$

From this it is clear that:

$$\text{RSS}(y, X, \beta) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The Maximum Likelihood Estimates (MLE) for $(\beta, \sigma_\epsilon^2)$ can then be derived by minimizing the negative log-likelihood:

$$\{\boldsymbol{\beta}, \sigma_\varepsilon^2\} \stackrel{\mathrm{m}}{=} \frac{n}{2}\log(\sigma_\varepsilon^2) + \frac{1}{2\sigma_\varepsilon^2} RSS(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})$$

Which shows:

$$MLE(\boldsymbol{\beta}) = \widehat{\boldsymbol{\beta}}_{OLS}$$
$$MLE\left(\sigma_\varepsilon^2\right) = \widehat{\sigma_\varepsilon^2} = \frac{RSS\left(y, X, \widehat{\beta}_{OLS}\right)}{\mathrm{n}}$$

and that

$$\{\boldsymbol{\beta}, \sigma_\varepsilon^2\} \overset{\mathrm{m}}{=} \frac{n}{2}\log(\sigma_\varepsilon^2) + \frac{1}{2\sigma_\varepsilon^2} RSS(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})$$

Often, researchers use IID normal priors for β, but, since there is no prior knowledge uninformed / flat priors were used.

From this, one can draw the conclusion that, assuming a flat prior as mentioned above, the resulting posterior will be:

$$p(\beta_1, \ldots, \beta_n | y_1, \ldots, y_n) \propto 1 \times \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \beta_1 - \beta_2 x_i - \ldots - \beta_p x_i\right)^2\right\}$$

which:

$$e^{-\frac{1}{2\sigma^2}\Sigma\left(y_i - \beta - \beta x_i - \ldots - \beta_p x_i\right)^2}$$

Specifically, for our regression function our:

Likelihood function:

Likelihood Function of Undercount (Y):

$$p\left(\mathrm{UNDERCOUNT\ (Y)} | \mathrm{EQUIP, ECON, GORE, BUSH, BALLOTS}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathrm{UNDERCOUNT}_i - \beta_1 - \mathrm{EQUIP}_2 x_i - \mathrm{ECON}_3 x_i - \mathrm{GORE}_4 x_i - \mathrm{BUSH}_5 x_i - \mathrm{BALLOTS}_6 x_i\right)^2\right\}$$

Prior:

Flat Prior:
$$p(\beta_1, \ldots, \beta_p) \propto 1$$

Posterior Distributions:

Posterior Distribution of Regression Coefficients:

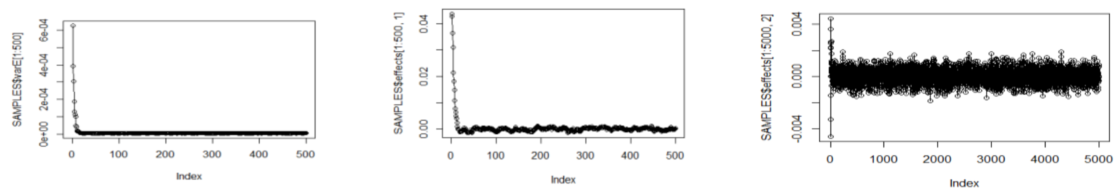
$$p\left(\mathrm{EQUIP, ECON, GORE, BUSH, BALLOTS} | \mathrm{UNDERCOUNT\ (Y)}\right) \propto 1 \times \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathrm{UNDERCOUNT}_i - \beta_1 - \mathrm{EQUIP}_2 x_i - \mathrm{ECON}_3 x_i - \mathrm{GORE}_4 x_i - \mathrm{BUSH}_5 x_i - \mathrm{BALLOTS}_6 x_i\right)^2\right\}$$

In the generic formed a posterior is defined in the form of:

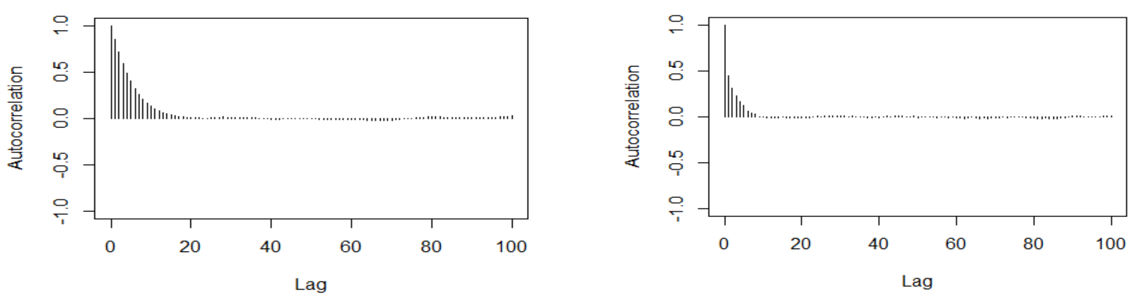
$$p(\theta|y) \propto p(\theta) * p(y|\theta)$$

The GibbsMLR function was then run and it created 15,000 samples using this function.

One can then conduct Post-Gibbs analysis (trace plot, autocorrelation, decide on burn-in and thinning, provide posterior means, posterior SDs and posterior credibility regions, estimate and report MC error).



From the above plots it was clear 500 iterations should be discarded from the data as burn-in order to achieve stationarity. Since 500 is the largest amount of burn-in required that is the value that was removed from the data. Autocorrelation plots were then examined in order to perform thinning.



In both instances the autocorrelation approaches 0 as Lag increases to 20 Thus 1 value should be sampled every 20 indexes to control for autocorrelation. Thus, one should resample with value 20,40, …. to n hundred (end of our data set). This process is called thinning our data. This was then applied to our Gibbs Sampler function.

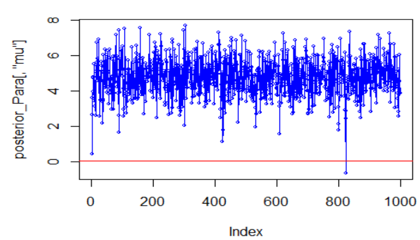
Now, both burn-in and thinning of the data, prerequisites for Bayesian work, had been completed, an HPD interval for all the variables was then created.

```
                   lower        upper
(Intercept) -1.020122e-03 1.054924e-03
equipOS-CC  -1.054852e-03 9.970905e-04
equipOS-PC  -1.455529e-03 1.347037e-03
equipPAPER  -3.690067e-03 3.686867e-03
equipPUNCH  -1.464422e-03 1.475174e-03
econpoor    -9.557012e-04 9.943387e-04
econrich    -1.599204e-03 1.585635e-03
gore        -1.866380e-07 2.764236e-07
bush        -1.706217e-07 2.501421e-07
ballots     -2.322399e-07 1.739616e-07
undercount   9.812473e-01 1.020100e+00
attr(,"Probability")
[1] 0.95
```

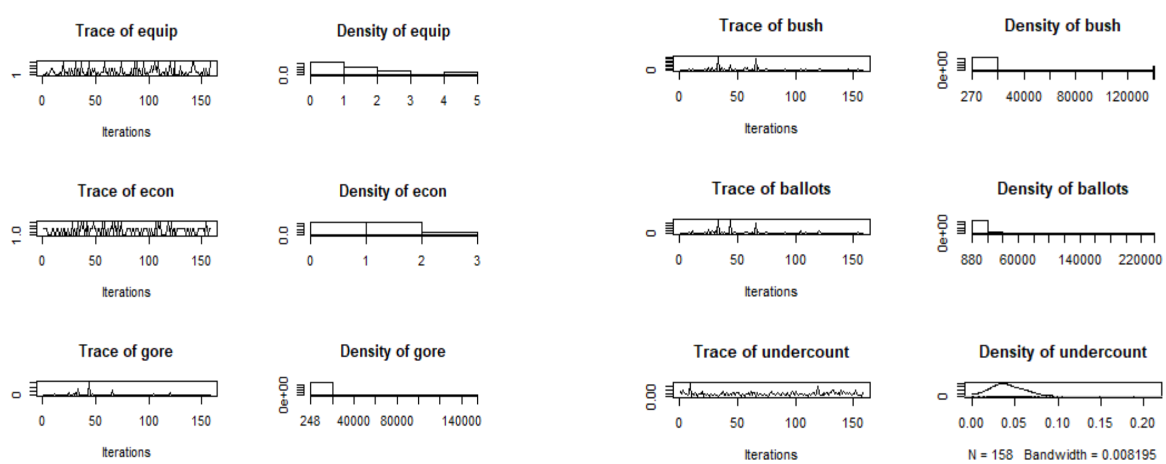
```
Iterations = 1:14500
Thinning interval = 1
Number of chains = 1
Sample size per chain = 14500

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

      Mean            SD        Naive SE Time-series SE
   6.614e-06     7.669e-07     6.369e-09      6.895e-09

2. Quantiles for each variable:

     2.5%       25%       50%       75%     97.5%
5.292e-06 6.077e-06 6.556e-06 7.085e-06 8.276e-06
```

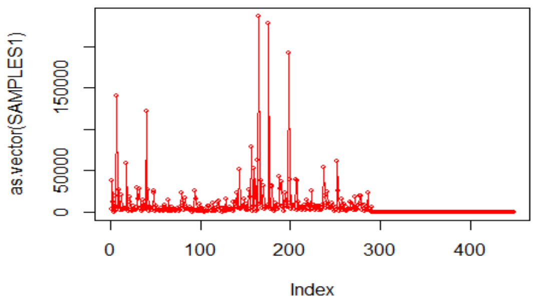
```
Iterations = 1:14500
Thinning interval = 1
Number of chains = 1
Sample size per chain = 14500

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                  Mean        SD  Naive SE Time-series SE
(Intercept) -1.122e-05 5.246e-04 4.357e-06      1.426e-05
equipOS-CC  -1.060e-05 5.256e-04 4.365e-06      8.801e-06
equipOS-PC  -2.306e-05 7.112e-04 5.906e-06      1.141e-05
equipPAPER   3.617e-06 1.881e-03 1.562e-05      1.627e-05
equipPUNCH  -5.692e-06 7.550e-04 6.270e-06      9.919e-06
econpoor    -1.347e-05 4.987e-04 4.141e-06      9.311e-06
econrich     1.096e-06 8.151e-04 6.769e-06      1.500e-05
gore         3.773e-08 1.256e-07 1.043e-09      2.638e-08
bush         3.543e-08 1.157e-07 9.609e-10      2.440e-08
ballots     -3.429e-08 1.115e-07 9.256e-10      2.927e-08
undercount   1.001e+00 9.823e-03 8.158e-05      2.665e-04

2. Quantiles for each variable:

                 2.5%        25%        50%       75%     97.5%
(Intercept) -1.045e-03 -3.607e-04 -1.970e-05 3.358e-04 1.042e-03
equipOS-CC  -1.033e-03 -3.663e-04 -8.718e-06 3.439e-04 1.028e-03
equipOS-PC  -1.438e-03 -5.035e-04 -2.424e-05 4.549e-04 1.371e-03
equipPAPER  -3.667e-03 -1.274e-03 -2.948e-07 1.253e-03 3.716e-03
equipPUNCH  -1.487e-03 -5.183e-04 -8.784e-07 5.027e-04 1.458e-03
econpoor    -9.916e-04 -3.533e-04 -1.061e-05 3.236e-04 9.673e-04
econrich    -1.594e-03 -5.443e-04  9.642e-06 5.426e-04 1.595e-03
gore        -1.759e-07 -6.355e-08  3.553e-08 1.299e-07 2.928e-07
bush        -1.607e-07 -5.594e-08  3.001e-08 1.218e-07 2.636e-07
ballots     -2.605e-07 -1.178e-07 -3.118e-08 5.691e-08 1.533e-07
undercount   9.810e-01  9.938e-01  1.001e+00 1.007e+00 1.020e+00
```

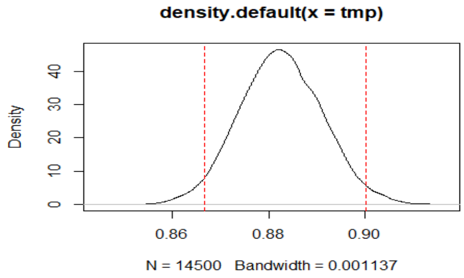
Now that the posterior is derived Gibbs-Sampling was performed again to plot the posterior distribution of undercount.



The distribution of our samples at the variable level can also be examined:







Now one can draw from our prior and perform hypothesis testing. This is not effective as with the prior chosen as there is nothing to test against. The density of our posterior * sample of the prior can be examined though.

It can be seen from this regression output that the Bayesian Normal Model modeled the data similarly to the output of the lm() function used above. The Gibbs Sampler expanded the data set, which decreased cross-variable correlation and co-linearity significantly. These aspects of the Bayesian Analysis are strengths when compared to the linear model.

Conclusion and Results

In the context of our problem, the result is clear. The level of income had a statistically significant effect on whether a location was more likely to be undercounted. This was surprisingly uncorrelated with any one type of a machine (at a significant level) which showed that economic status was a better predictor than voting machine. This is likely because in many instances the lowest performing voting machines are already in lower economic areas (resulting in co-linearity of these variables).

Overall, although this relationship was likely not very linear both models were able to successfully examine the relationship between the predictors and the predicted undercount.

Disscussions of Strengths and Weaknesses of the Models (Bayesian vs. Frequentist)

The strengths of the frequentist model are its' interpretabilitiy. We were able to directly examine the effects of coefficients and add and remove variables to examine different relationships.

The weaknesses are that it is limited to modeling linear relationships, of which the data likely is not (from the results of the regression diagnostics).

The Bayesian Model creates a more effective data set, implementing stationarity and reducing autocorrelation. Its' main weakness is in interpretation. Though we have intervals and means (and standard errors) for an expanded data set we do not have that easily interpretable function that a reader can immediately understand and apply to make inference on new situations. This is especially a weakness in the social sciences within which researchers often would like to extrapolate models to explain more than one unique situation.

Overall both models did a relatively good job modeling the data and creating the ability to predict undercount at the county level.