STAT 4444 Final Project

Ryan Webster Class ID: 76

Introduction

This paper will employ Multiple Linear Regression from both frequentist and Bayesian perspectives to develop a linear model aimed at predicting and identifying key factors influencing President Obama's vote share in North Carolina during the 2012 election. These MLR models will utilize various race, economic, and educational demographic factors from each of the 100 counties in North Carolina with the response variable being President Obama's vote share in the respective county.

Data

In this dataset, there are 100 rows, one for each of the 100 counties in North Carolina along with 12 covariates containing the demographic data of its respective county. Shown below in *Table 1* is a data dictionary of each of the variables in the dataset.

Table 1: Data Dictionary of NC Dataset

Variable	Description
County	The name of the county
PCTOBAMA	Obama's vote share
Pop2010	Population as of 2010
PopPerSqMile	Population per square mile
PctWhite	Percent of population that is White
PctBlack	Percent of population that is Black
PctAmericanIndian	Percent of population that is American Indian
PctAsian	Percent of population that is Asian
PctHisp	Percent of population that is Hispanic
DiversityIndex	Value which represents how 'diverse' a county is.
pct_Uninsured	Percent of population that is uninsured
post_sec_edu	Percent of population with post-secondary education

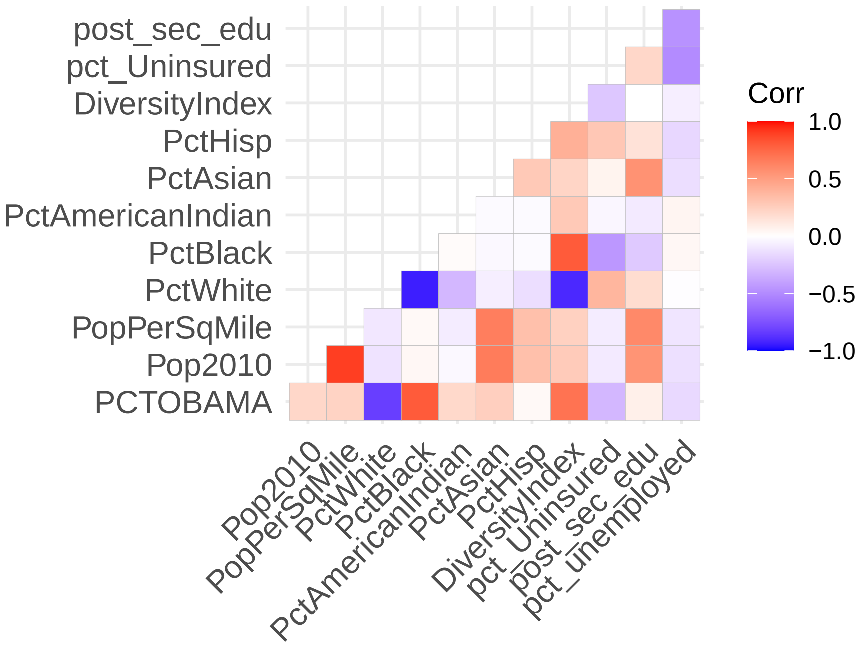
Variable	Description
pct_unemployed	Percent of population unemployed

Shown below in *Table 2* are the summary statistics of the dataset:

Table 2: Summary Statistics

	Mean	Median	SD	Min	Max
PCTOBAMA	0.44	0.42	0.12	0.24	0.76
Pop2010	95354.83	55621.50	141743.07	4407.00	919628.00
PopPerSqMile	195.44	113.05	260.39	9.50	1755.50
PctWhite	0.71	0.74	0.18	0.29	0.97
PctBlack	0.21	0.18	0.17	0.00	0.63
PctAmericanIndian	0.02	0.00	0.05	0.00	0.38
PctAsian	0.01	0.01	0.01	0.00	0.07
PctHisp	0.07	0.06	0.04	0.01	0.21
DiversityIndex	41.86	46.90	16.67	9.70	72.60
pct_Uninsured	21.26	21.00	3.81	12.00	30.00
post_sec_edu	51.80	50.75	9.40	29.20	76.20
pct_unemployed	11.06	11.00	2.18	6.60	16.10

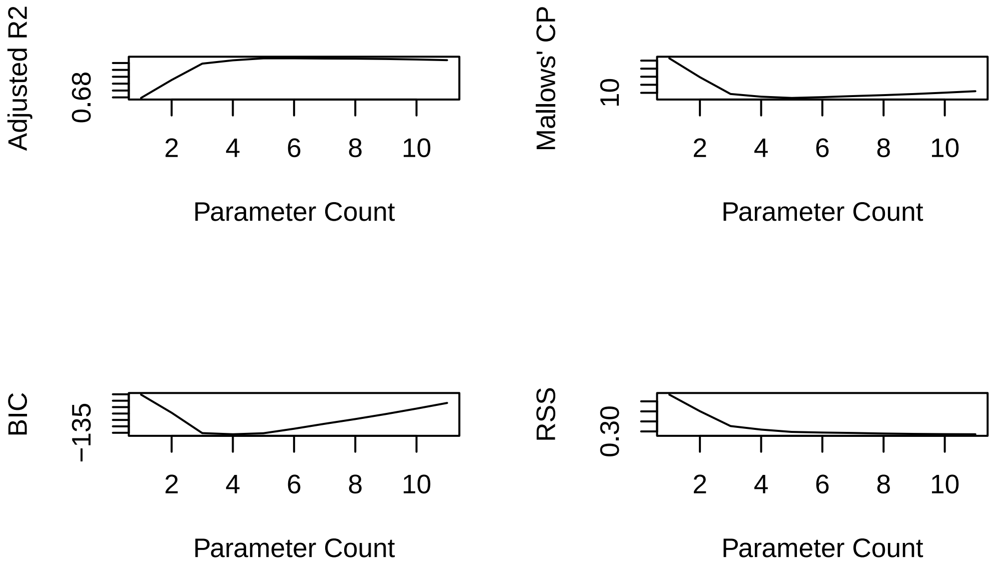
To begin the MLR process, it's necessary to identify linear trends between the predictor (PCTOBAMA) to each of the other factors. A matrix scatter plot should be avoided due to the large number of parameters, which would make the plots unreadable. Instead, a correlation matrix heatmap was made which measures variables' correlation coefficient between each other.



From observing this correlation heatmap, there seems to be a strong linear trend between PCTOBAMA and PctWhite, PctBlack, and DiversityIndex. More specifically, a negative correlation between PCTOBAMA and PctWhite and a positive correlation between PCTOBAMA and PctBlack and DiversityIndex. These results makes sense considering how different races vote is well knows. Typically, at least in 2012, white voters typically favored the Republican candidate while non-white voters favored the Democrat candidate. This seems to not be an exception in North Carolina during the 2012 election. However, using the variables PctWhite, PctBlack, and Diversity Index as factors in the analysis could cause issues concerning multicollinearity as these variables would have high linear dependence between them (PctWhite is affected by PctBlack and vice versa. Additionally, Diversity Index most likely is influenced by PctWhite and PctBlack). Also, it's possible there are variables that aren't statistically significant in predicting PCTOBAMA, but could be if combined with other variables. So, a best subset model selection method will be used to determine what factors to use.

Model Selection

Using the leaps library, a best subset selection method was ran, which tracks the best model of each number of parameters based on various metric performances. Shown below are the model performance based on the best model for each number of parameters for Adjusted R^2, Mallows' CP, BIC, and RSS.



From these results, it could be concluded that the best and simplest model would be the one with three variables. However, because this result contains variables PctWhite and DiversityIndex, there is high concern for multicolinearity, which can cause serious consequences to our interpretation of the model results. This can be concluded by looking at their VIF values.

Table 3: Multicolinearity Check

	VIF
PctWhite	7.036971
DiversityIndex	6.802026
post_sec_edu	1.234614

With VIF values greater than 5, it can be concluded that the multicollinearity issues between PctWhite and DiversityIndex are valid. Some possible solutions could be to use LASSO or Ridge regression to shrink or zero out coefficients, but instead the best model with two variables, PctWhite and post_sec_edu will be used instead.

MLR: Frequentist

Using the lm() function, a MLR model will be created as PCTOBAMA as the response and PctWhite and post_sec_edu as the covariates.

Table 4: Coefficient Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7118275	0.0412674	17.24915	0.00e+00
PctWhite	-0.6060566	0.0370054	-16.37754	0.00e+00
post_sec_edu	0.0031265	0.0006993	4.47094	2.11e-05

According to the model, parameters PctWhite and post_sec_edu were significant with p-values < 0.05. This is despite the fact that the relationship between post_sec_edu and PCTOBAMA was not linear, but together with PctWhite, it is. As one would expect, the higher the share of the White population in a county negatively impacts President Obama's vote share. Additionally, counties with higher percent of post-secondary education contributed to Obama's vote share, albeit a very small amount, but statistically significant.. The MLR equation for predicting President Obama's vote share by county follows:

$$PCTOBAMA = 0.712 - 0.606(PctWhite) + 0.003(post\_sec\_edu) + \epsilon$$

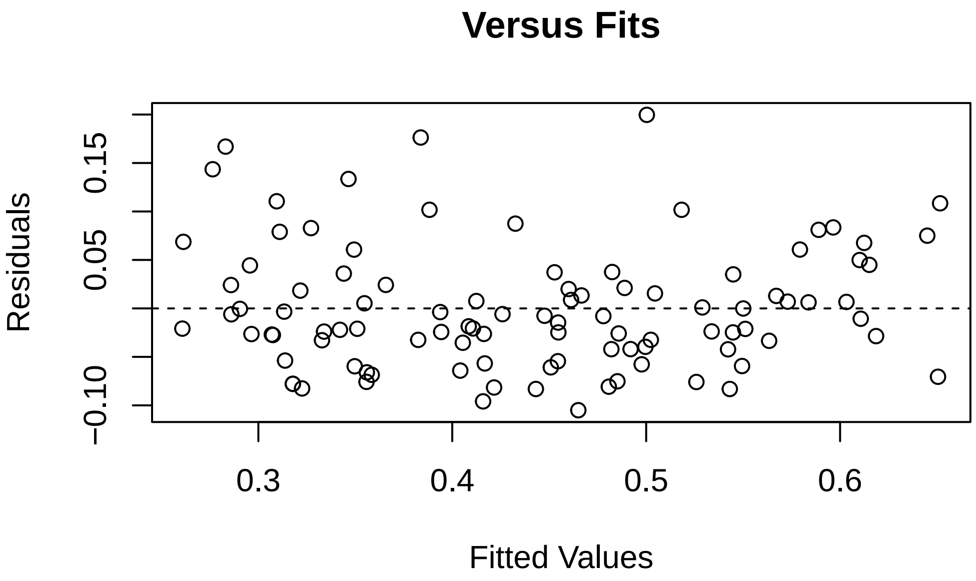
So, starting at 71.2% for President Obama in a North Carolina county, subtract 0.606 for every percent of White people, assuming post secondary education is constant, and add 0.003 for every percent of post secondary education in a county, assuming percent of White people is constant.

Assumptions

In order for the MLR to produce valid results, there are three assumptions that must be met: constant variance, normally distributed errors, and independent observations.

Constant Variance

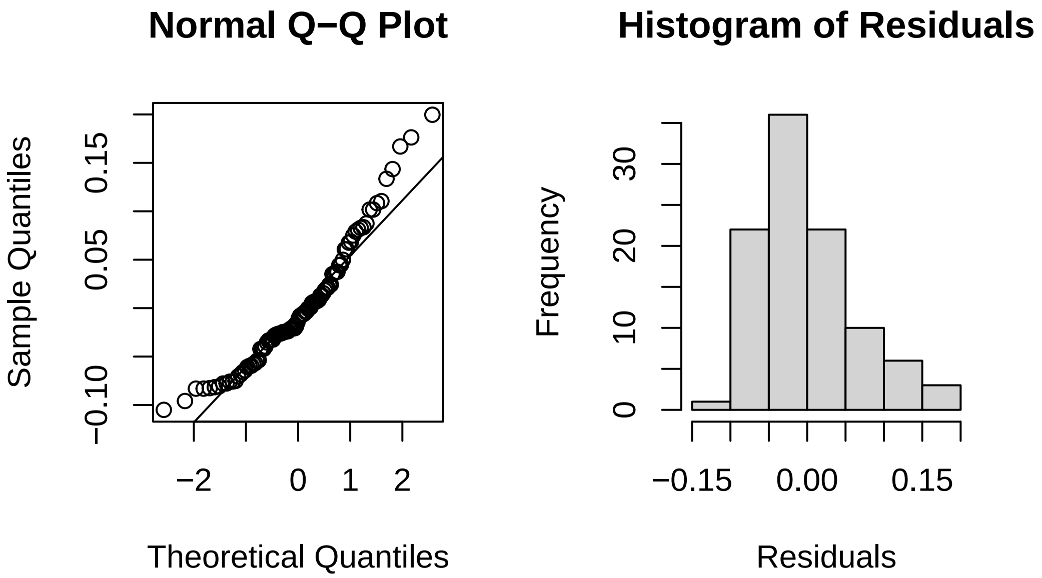
To test for constant variance, a scatterplot showing the fitted values versus the residuals should be made, ideally with no obvious patterns and difference of spread along the horizontal line.



The assumption of constant variance passes as there is no sign of patterns or difference in spread in the fitted values vs. residuals scatter plot.

Normally Distributed Errors

Testing for Normally distributed errors is done by creating a Q-Q plot. Ideally, the points should lie on the 45 degree line.



There seems to be some concern for heavy tails and right skewness. To further verify this conclusion, the Shapiro-Wilk test will be used as a statistical hypothesis test where the null hypothesis is that the errors are normally distributed and the alternative is that the errors are not normally distributed.

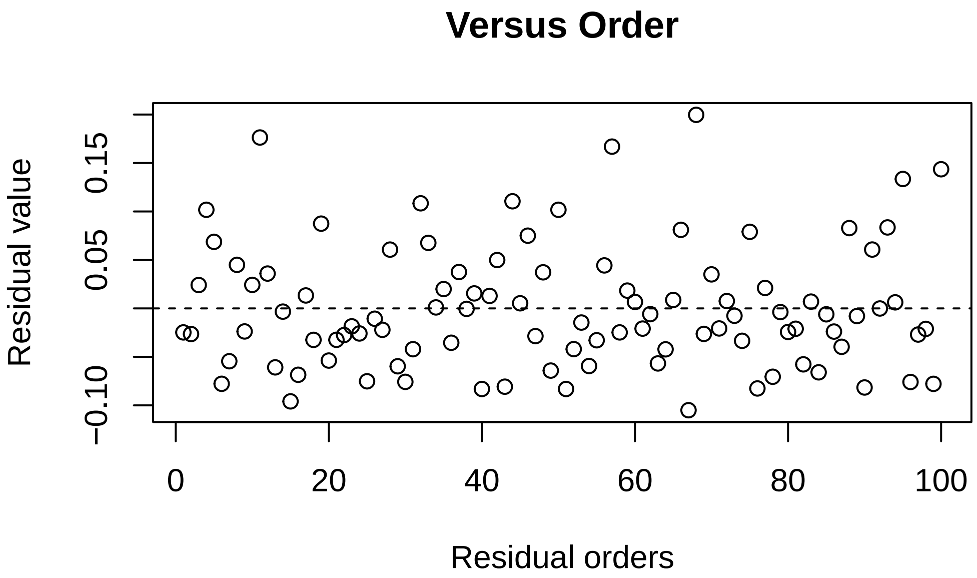
```
	Shapiro-Wilk normality test

data:  mlr_freq$residuals
W = 0.94098, p-value = 0.0002214
```

This result further verify's the concern for non-normal residuals. We will have to be sure to note this break in our assumptions in the analysis.

Independence

To test for independence, a scatterplot will be made to measure the order of the residual and its actual value. Ideally, having no obvious pattern along the horizontal line is a good sign of independence. This shouldn't be an issue as this isn't a time series data set and the result in one county wouldn't influence the other.



Like expected, there is no obvious pattern. A statistical test called the Durbin Watson test can be used to test for linearity/autocorrelation where the null hypothesis is the data is independent and the alternative is that the data isn't independent.

```
 lag Autocorrelation D-W Statistic p-value
   1      -0.1789312      2.304947   0.106
 Alternative hypothesis: rho != 0
```

The result of the Durbin Watson test concludes that the data is independent by not rejecting the null hypothesis.

MLR: Bayesian

To begin the Bayesian analysis, the rjags library will be used to as the Gibbs' Sampler to estimate coefficient posterior distributions for β_0, β_1, β_2, and σ^2.

Priors

For the priors of each of the parameters, an uninformative prior will be used on β_1, β_2, and σ^2 while an informative prior can be found for the intercept based on the hypothetical question of how much vote share would Obama win in a county that is 0% White and 0% post-secondary education, as the intercept of the MLR is simply the expected value when all other values are zeroed out. Although this scenario is unrealeastic, one could expect this hypothetical county would give Obama a high vote share of around 90% based on vote shares with similar demographics as this hypothetical (i.e., Claiborne County, Mississippi, where Obama won 88% of the vote (2012) in this 14% White county).

β_1: PctWhite, β_2: post_sec_edu

$$\beta_0 \sim N(0.9, 0.05) \quad \beta_1 \sim N(0, 100) \quad \beta_2 \sim N(0, 100) \quad \sigma^2 \sim IG(0.1, 0.1)$$

Gibbs' Sampler

Table 5: Coefficient Summary

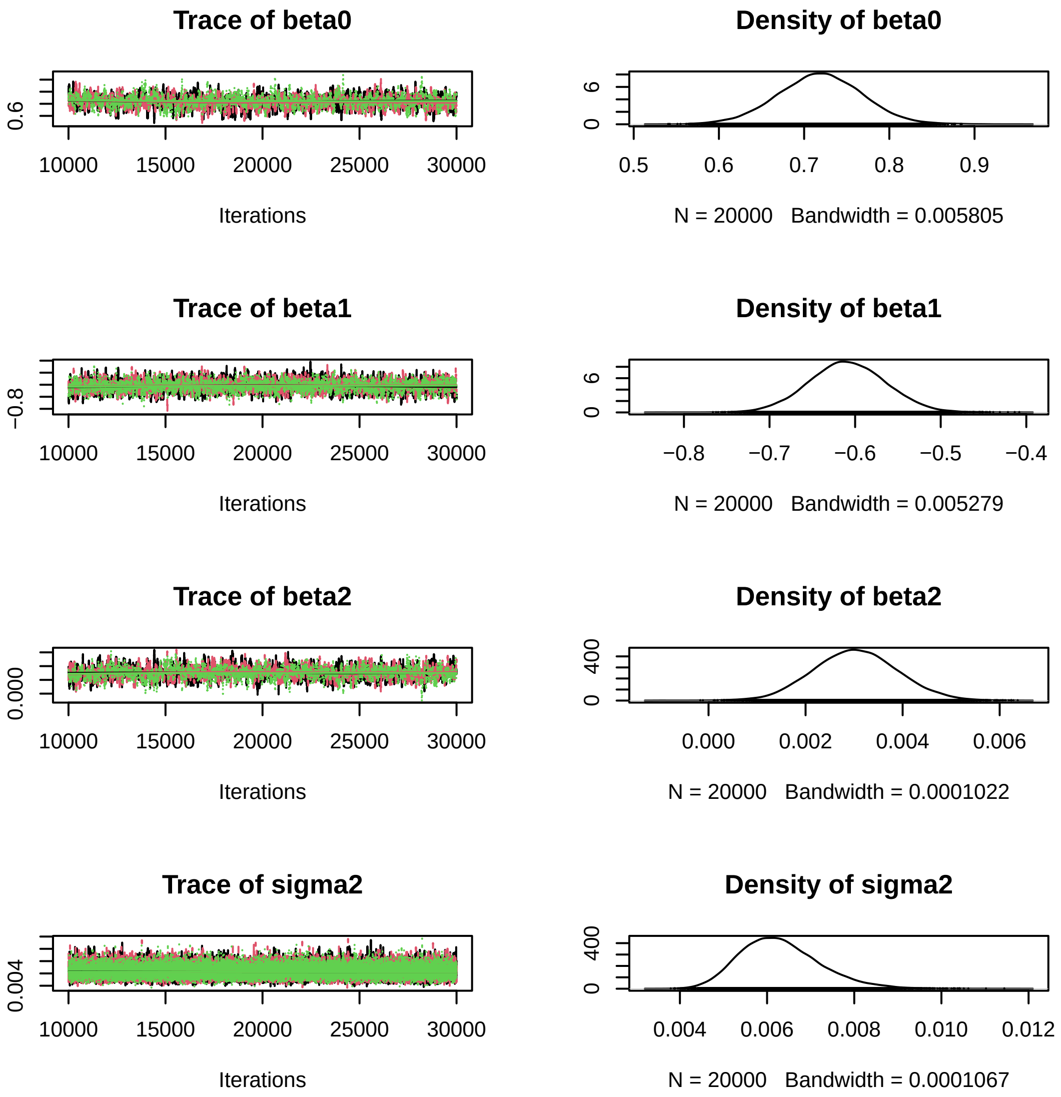
	Mean	SD	Naive SE	Time-series SE
beta0	0.7188239	0.0498176	0.0002034	0.0017819
beta1	-0.6088125	0.0449686	0.0001836	0.0010585
beta2	0.0030333	0.0008710	0.0000036	0.0000287
sigma2	0.0063097	0.0009248	0.0000038	0.0000043

Table 6: Coefficient Quantiles

	2.5%	25%	50%	75%	97.5%
beta0	0.6204211	0.6856494	0.7188903	0.7519047	0.8166436
beta1	-0.6964686	-0.6390527	-0.6094018	-0.5787493	-0.5201651
beta2	0.0013598	0.0024453	0.0030258	0.0036112	0.0047647
sigma2	0.0047571	0.0056519	0.0062231	0.0068700	0.0083749

Trace Plots and Parameter Posterior Distribution

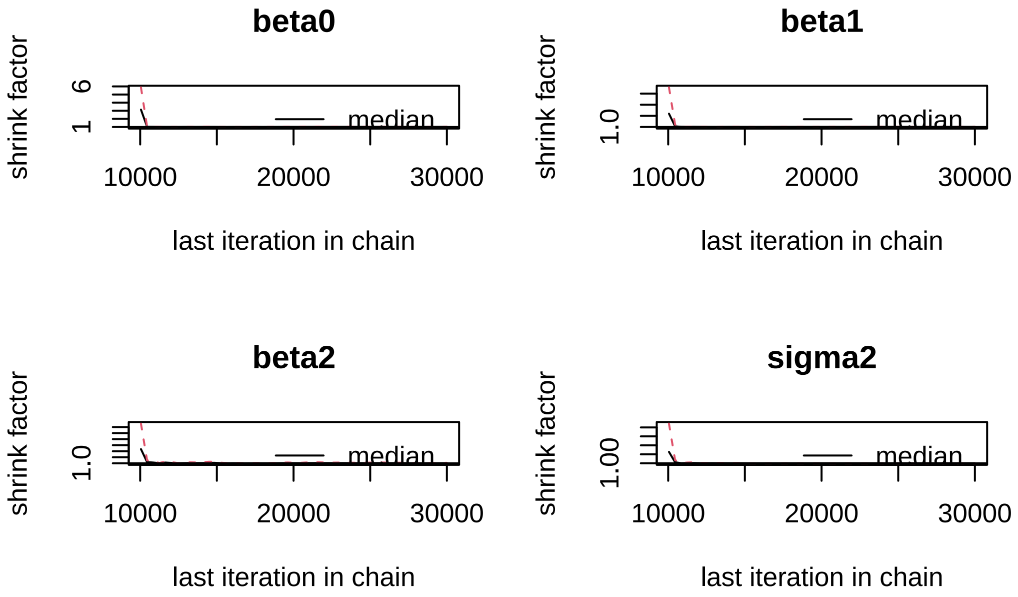
Shown below are the trace plots and density distributions of each of the parameters:



In this example, three chains were used and they all seemed to behave well according to the trace plots.

Gelman Plots and Effective Sample Size

To determine if the parameters converged, Gelman Plots can be used to determine if the effects of the initial value has successfully been removed by observing the variances between the different chains.



According to the Gelman plots, each of the parameters converged as the median values quickly approached zero. Now, to observe the sample size of the actual independent information in the posterior distribution for each parameter, the effectiveSize function can be used to calculate those values.

Table 7: Effective Sample Size

	Effective Sample Size
beta0	778.4654
beta1	1805.0799
beta2	918.9869
sigma2	45911.4489

The effective sample sizes are sufficient for this model.

Comparisons

Overall, the coefficients between the frequentest, Ordinary Least Squares approach should be similar to the Bayesian approach, especially as the priors for β_1, β_2, and σ^2 were all uninformative. Shown below is a table comparing the coefficients computed from the OLS method and the mean coefficient values of the respective posterior distributions:

Table 8: MLR Coefficient Values of OLS and Bayesian

	Frequentist	Bayesian
(Intercept)	0.7118275	0.7188239
PctWhite	-0.6060566	-0.6088125
post_sec_edu	0.0031265	0.0030333

As expected, the coefficient values from OLS and Bayesian are nearly identical. This is due to the priors of the coefficients being uninformed (beside the intercept, which its prior mean was greater than the OLS estimate, so the Bayesian estimate of the intercept is greater than the OLS).