Project Paper 1 - Revised

Rudi Herrig and Jordan Kim

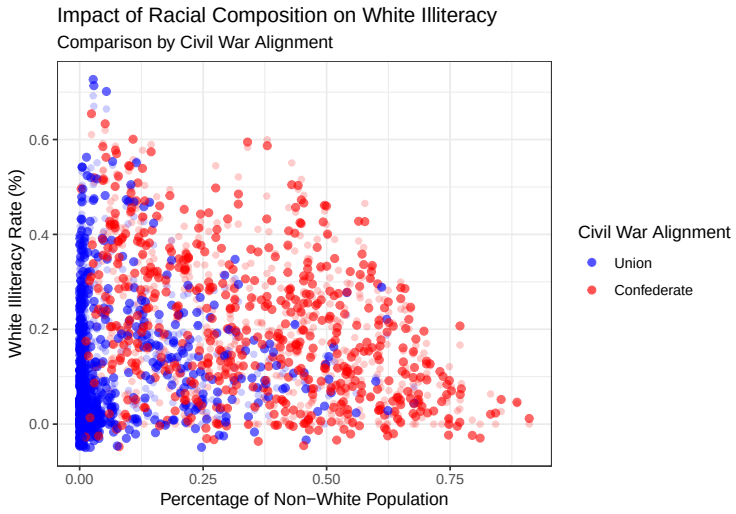
5 May 2024

I wanted to try creating a binary response variable indicating whether a county joined the South because the state in which it is located fought against the Union. I will use the following states: Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, and Virginia. I will create a new variable, Confederate, which is equal to 1 if the county is in one of these states, and 0 if the county is not in one of these states. I will use the mutate function to create this new variable.

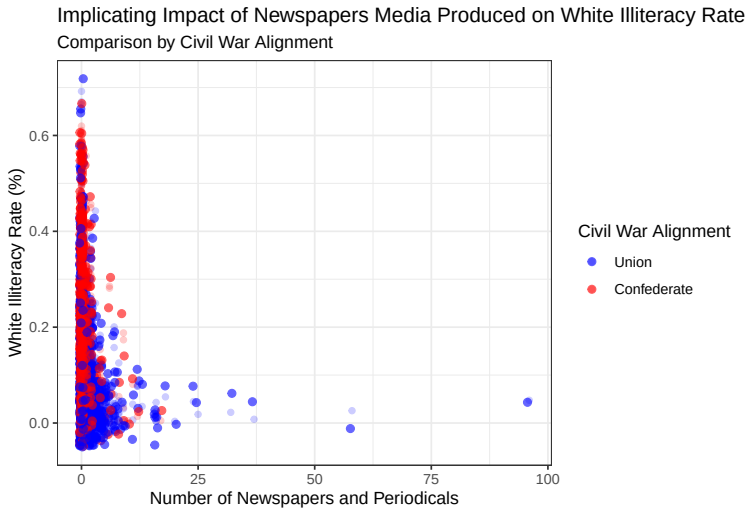
Filtering for only the columns we're interested in and creating a new list `WhiteIlliterate`

Visualizing data we're interested in to learn more about relationships among data set with Blue representing Union counties and Red representing Confederate counties

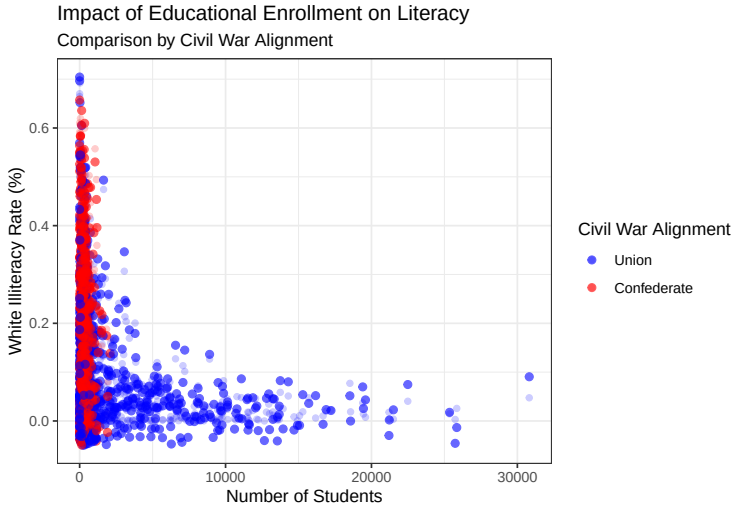
The visualization of the response variable, White Illiteracy, with Percentage of Non-White Population as potential predictor



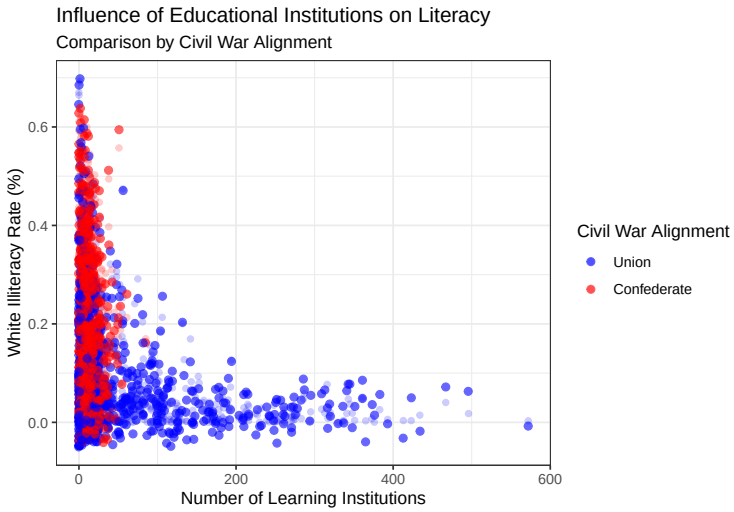
Visualizing the response variable, White Illiteracy, with Number of Newspapers and Periodicals produced as potential predictor



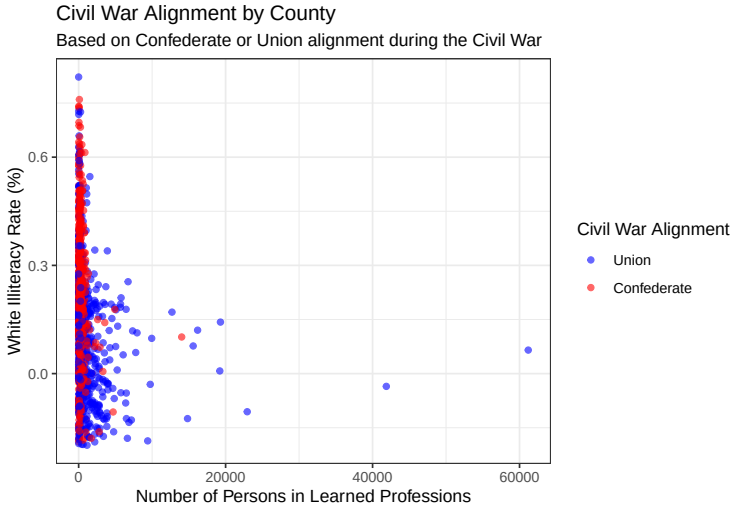
Visualizing the response variable, White Illiteracy, with Number of Students as potential predictor



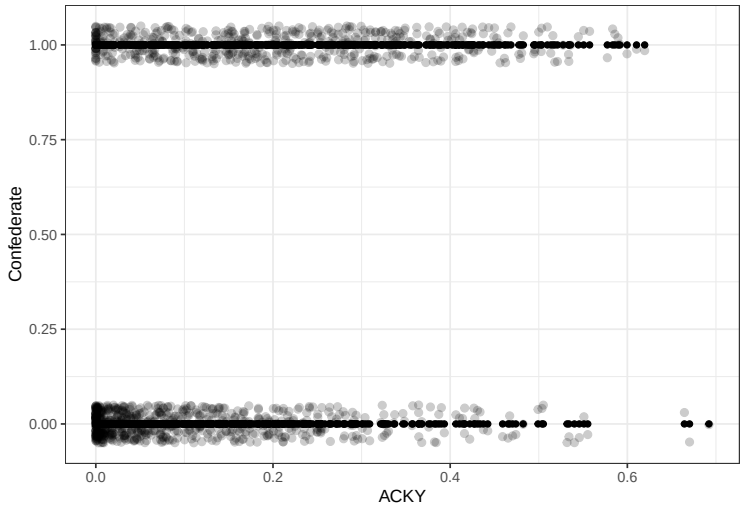
Visualizing the relationship between the Number of Learning Institutions and White Illiteracy



Visualizing the response variable, White Illiteracy, with Number of Persons in Learned Professions Requiring Literacy as potential predictor



Visualizing new binary variable, `Confederate`, as potential response, with White Illiteracy as potential predictor



Null Model

```
## [1] 0.149582
```

The mean White Illiteracy Rate for counties is 0.149582, or 14.96%. We will use the RMSE to evaluate our models' performances. We will also look at the distributions of the predictor variables and the correlation between the predictor variables and the response variable.

We will also utilize various techniques to try to find the best model, from shrinkage, dimension reduction, model selection and variable selection. We will also look at the distributions of the predictor variables and the correlation between the predictor variables and the response variable.

Splitting Data into Training and Testing Sets

Forward Stepwise Variable Selection

```
## Start:  AIC=-2429.75
## ACKY ~ (STATE + COUNTY + ACD001 + ACN001 + ACY + ACZ + `Total - ACZ` +
##     ACK001 + ACY005 + ACY006 + ACY007 + ACY008 + ACY009 + ACY010 +
##     ACY011 + ACY012 + ACY013 + ACY018 + ACY019 + ACY020 + ACY021 +
##     ACY022 + ACY023 + ACY024 + ACY025 + ACY026 + ACZ001 + ACZ002 +
##     ACZ003 + ACZ004 + ACZ005 + ACZ006 + ACZ007 + ACZ008 + ACZ009 +
##     ACZ010 + ACZ011 + ACZ012 + ACZ013 + ACZ014 + ACZ015 + ACZ016 +
##     ACZ017 + ACZ018 + ACZ019 + ACZ020 + ACZ021 + ACZ022 + ACZ023 +
##     ACZ024 + AC1001 + AC1002 + AC1003 + AC1004 + AC1005 + AC1006 +
##     AC1007 + AC3001 + AC3002 + AC3003 + ACE001 + ACE002 + ACF001 +
##     ACF002 + ACF003 + ACF004 + ACG001 + ACH001 + ACH002 + ACH003 +
##     ACI001 + ACI002 + ACI003 + ACJ001 + ACL001 + ACL002 + ACM001 +
##     ACM002 + ACM003 + ACM004 + ACO001 + ACP001 + ACQ001 + ACR001 +
##     ACR002 + ACS001 + ACS002 + ACS003 + ACT001 + ACT002 + ACU001 +
```

```
##     ACU002 + ACU003 + ACU004 + ACV001 + ACV002 + ACX001 + ADB001 +
##     ADB002 + ADB003 + ADB004 + ADB005 + ADB006 + ADB007 + ADB008 +
##     ADB009 + ADB010 + ADB011 + ADB012 + ADB013 + ADB014 + ADB015 +
##     ADB016 + ADB017 + ADB018 + ADB019 + ADB020 + ADB021 + ADB022 +
##     ADB023 + ADB024 + ADB025 + ADB026 + ADB027 + ADB028 + ADB029 +
##     ADB030 + ADE001 + ADE002 + ADE003 + ADE004 + ADE005 + ADE006 +
##     ADE007 + ADE008 + ADE009 + ADE010 + ADE011 + ADE012 + ADE013 +
##     ADE014 + ADE015 + ADE016 + ADE017 + ADE018 + ADE019 + ADE020 +
##     ADE021 + ADE022 + ADE023 + ADE024 + ADE025 + ADE026 + ADE027 +
##     ADE028 + ADE029 + ADE030 + ADE031 + ADE032 + ADE033 + ADE034 +
##     ADE035 + ADE036 + ADC001 + ADC002 + ADC003 + ADC004 + ADC005 +
##     ADC006 + ADC007 + ADC008 + ADC009 + ADC010 + ADC011 + ADC012 +
##     ADC013 + ADC014 + ADC015 + ADC016 + ADC017 + ADC018 + ADC019 +
##     ADC020 + ADC021 + ADC022 + ADC023 + ADC024 + ADC025 + ADC026 +
##     ADC027 + ADC028 + ADC029 + ADC030 + ADC031 + ADC032 + ADC033 +
##     ADC034 + ADC035 + ADC036 + ADC037 + ADC038 + ADC039 + ADD001 +
##     ADD002 + ADD003 + ADD004 + ADD005 + ADD006 + ADD007 + ACZYZ +
##     ACH + ACI + ACL + ACM + AC1 + ACE + ACS + ACSW + ACSNW +
##     ACSP + ACSWP + Confederate) - COUNTY
```

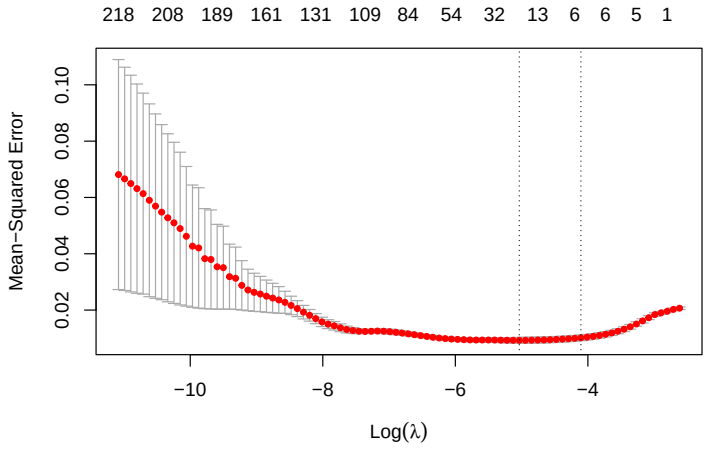
```
##                                Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)                2.753008e-01 2.118405e-02 12.995663  8.571948e-36
## STATEDelaware             -3.160336e-01 9.549080e-02 -3.309572  9.677514e-04
## STATEFlorida Territory    -8.034411e-02 2.651708e-02 -3.029900  2.509235e-03
## STATEIllinois             -1.373486e-01 2.215019e-02 -6.200788  8.207424e-10
## STATEIndiana              -1.502727e-01 2.215345e-02 -6.783263  2.007867e-11
## STATEIowa Territory       -1.995833e-01 2.873682e-02 -6.945213  6.796420e-12
## STATEMaine                -2.001157e-01 4.645661e-02 -4.307584  1.813227e-05
## STATEMassachusetts        -3.188169e-01 8.216359e-02 -3.880270  1.111724e-04
## STATEMichigan             -1.926147e-01 2.611596e-02 -7.375365  3.435893e-13
## STATEMissouri             -5.085982e-02 1.915248e-02 -2.655521  8.044522e-03
## STATENew Hampshire        -2.038455e-01 5.497317e-02 -3.708090  2.203070e-04
## STATENew Jersey           -1.611524e-01 3.854395e-02 -4.181004  3.156604e-05
## STATENew York             -1.768204e-01 3.251438e-02 -5.438222  6.765137e-08
## STATEOhio                 -1.713742e-01 2.412095e-02 -7.104787  2.287149e-12
## STATEPennsylvania         -1.454449e-01 2.898873e-02 -5.017293  6.199474e-07
## STATERhode Island         -2.203985e-01 6.732461e-02 -3.273669  1.097996e-03
## STATEVermont              -2.480183e-01 3.754010e-02 -6.606756  6.374189e-11
## STATEVirginia             -8.025463e-02 1.842407e-02 -4.355965  1.461384e-05
## STATEWisconsin Territory -1.507911e-01 2.758868e-02 -5.465689  5.822654e-08
## ACK001                    2.396495e-04 7.819258e-06 30.648630 5.394789e-146
## ACF002                   -1.877310e-03 5.470887e-04 -3.431455  6.249158e-04
## ADB002                    4.779538e-06 1.833095e-06  2.607359  9.260246e-03
## ACZYZ                    -2.328188e-01 8.542861e-02 -2.725302  6.536436e-03
```

Forward Stepwise Selection using OLS led to `STATE(factors)` and `ACK001`, `ACF002`, `ADB002`, and `ACZYZ` as predictor variables. We will use these variables to build a multiple linear regression model.

```
## [1] 0.102337
```

And our Forward Stepwise selection method using AIC as criteria gave us a model using `STATE`, `ACK001`, `ACF002`, `ADB002`, and `ACZYZ` as predictor variables, with a testing RMSE of 0.102337**.**

Ridge and Lasso Regression (plus OLS glm)




```
## pdf
##   2
```

```
## [1] 0.006504086
```

```
## [1] 0.09876708
```

```
## [1] 0.9049026
```

```
## [1] 0.115614
```

Our glm model using Ordinary Least Squares model yielded RMSE of 0.9049026. Our model using Ridge Regression model yielded RMSE of 0.1075233. Our best model was using Lasso Regression model with testing RMSE of 0.1014527.

building full lasso model using all data

```
## [1] 0.0389958
```

```
##       STATEArkansas STATEIowa Territory       STATEKentucky       STATEMichigan
##        1.549911e-02       -3.103653e-04        2.457041e-02       -1.075394e-02
##       STATEMissouri       STATENew York STATENorth Carolina           STATEOhio
##        8.418677e-03       -6.833898e-03        8.451652e-03       -2.327616e-02
##   STATEPennsylvania              ACK001              ACZ014              AC1002
##       -1.111337e-02        1.740547e-04       -5.757079e-06       -1.883919e-06
##              ACM002              ACV001              ACX001              ADB028
##       -7.815777e-03       -5.073799e-06       -1.005754e-02       -2.482678e-04
```

```
##              ADE023              ADC016              ADC037              ADD002
##       -1.472832e-05        1.270062e-08       -5.139307e-08       -1.973548e-08
##         Confederate
##        3.899580e-02
```

```
## [1] "STATEArkansas"       "STATEIowa Territory" "STATEKentucky"
## [4] "STATEMichigan"       "STATEMissouri"       "STATENew York"
## [7] "STATENorth Carolina" "STATEOhio"           "STATEPennsylvania"
```

For ACKY as the response variable and also using STATE factors as part of the predictive model, we get STATES: "STATEArkansas" "STATEIowa Territory" "STATEKentucky" "STATEMichigan" "STATEMissouri" "STATENew York" "STATENorth Carolina" "STATEOhio" "STATEPennsylvania" as 'predictor variable factors(?)'.

```
##  [1] "ACK001"      "ACZ014"      "AC1002"      "ACM002"      "ACV001"
##  [6] "ACX001"      "ADB028"      "ADE023"      "ADC016"      "ADC037"
## [11] "ADD002"      "Confederate"
```

```
##                                                          ACK001
##             "Total White Persons Over 20 Who Cannot Read or Write"
##                                                          ACZ014
##                           "Slave >> Male >> 10 to 23 years of age"
##                                                          AC1002
##                                                    "Agriculture"
##                                                          ACM002
##                                                 "Weekly newspapers"
##                                                          ACV001
##                                                          "Male"
##                                                          ACX001
##                                                     "0=NO 1=YES"
##                                                          ADB028
##                           "Printing and Binding: Weekly newspapers"
##                                                          ADE023
##    "Manufacturing Establishments: Distilled and fermented liquors"
##                                                          ADC016
##                                "Manufacturing Establishments: Wool"
##                                                          ADC037
##                               "Manufacturing Establishments: Mills"
##                                                          ADD002
##                                   "Persons employed in agriculture"
##                                                           Confederate
## "Joined Confederacy and Raised Arms Against The United States = 1"
```

For ACKY as the response variable, Lasso gave us a model with the following predictor variables: - ACK001 - Total White Persons Over 20 Who Cannot Read or Write - ACZ014 - Slave » Male » 10 to 23 years of age - AC1002 - Persons Employed in Agriculture - ACM002 - Weekly newspapers - ACV001 - Total Male Population - ACX001 - 0=NO 1=YES (Navigable Waterway) - ADB028 - Printing and Binding: Weekly newspapers - ADE023 - Men Employed in Manufacturing Establishments: Distilled and fermented liquors - ADC016 - Capital Invested in Manufacturing Establishments: Wool - ADC037 - Capital Invested in Manufacturing Establishments: Mills - ADD002 - Persons employed in agriculture - Confederate - Joined Confederacy and Raised Arms Against The United States = 1, Union = 0 and as mentioned above, those predictor variables were in addition to the factor-predictor variable the STATE factors: "STATEArkansas" "STATEKentucky" "STATEMichigan" "STATEMissouri" "STATENew York" "STATENorth Carolina" "STATEOhio" "STATEPennsylvania"

Mistakenly performed modeling

For ACK001 as the response variable (mistakenly performed) Lasso gave us a model with the following variables: - ACD001 - Total Population - ACZ013 - Slave » Male » Under 10 years of age - AC1003 - Persons Employed in Commerce - AC1006 - Persons Employed in Navigation of canals, lakes, and rivers - AC3002 - Number of Deaf and Dumb White Persons 14 to 24 years old - ACE002 - Colored Blind Persons - ACF003 - Colored Insane and Idiot Persons » At public charge - ACF004 - Colored Insane and Idiot Persons » At private charge - ACH001 - Universities or colleges - ACM003 - Semi and tri-weekly newspapers - ACR002 - Total White Female Population - ACV002 - Total Female Population - ADB020 - Fisheries: Barrels of Pickled Fish - ADB026 - Bar Iron: Tons Produced - ADB029 - Printing and Binding: Semi- and tri-weekly newspapers - ADE002 - Men Employed in Mining Establishments: Lead - ADE005 - Men Employed in Mining Establishments: Anthracite coal - ADE017 - Men Employed in Manufacturing Establishments: Weapons - ADE019 - Men Employed in Manufacturing Establishments: Various metals - ADE024 - Men Employed in Manufacturing Establishments: Powder mills - ADC017 - Capital Invested in Manufacturing Establishments: Cotton - ADC036 - Capital Invested in Manufacturing Establishments: Carriages and wagons - ADD003 - Persons employed in commerce - ADD006 - Persons employed in inland navigation - ACH - Total Learning Institutions

```
## [1] 0.09226152
```

The full lasso model had an RMSE of 0.09297063 **using the best lambda value of** 0.007138227**. This is an improvement over the testing lasso RMSE of** 0.1075233**.**

For Paper 1 stuff

Histograms of predictor variables distributions