(Project Paper 1 - Revised +) Project Paper 2

Rudi Herrig and Jordan Kim

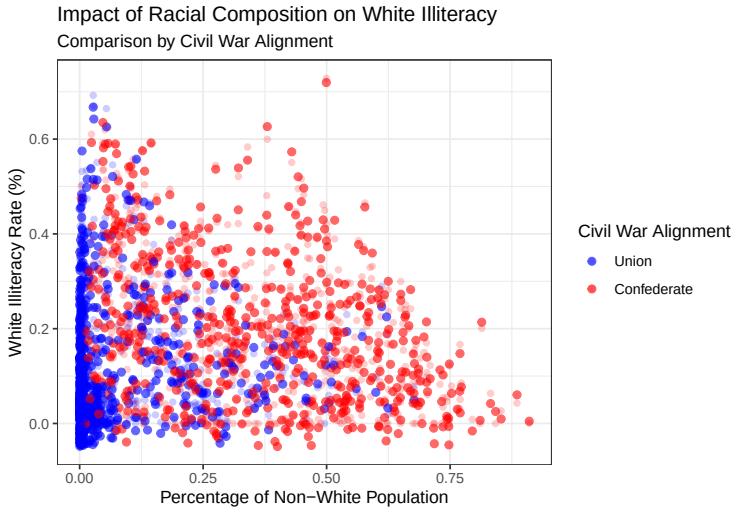
5 May 2024

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.

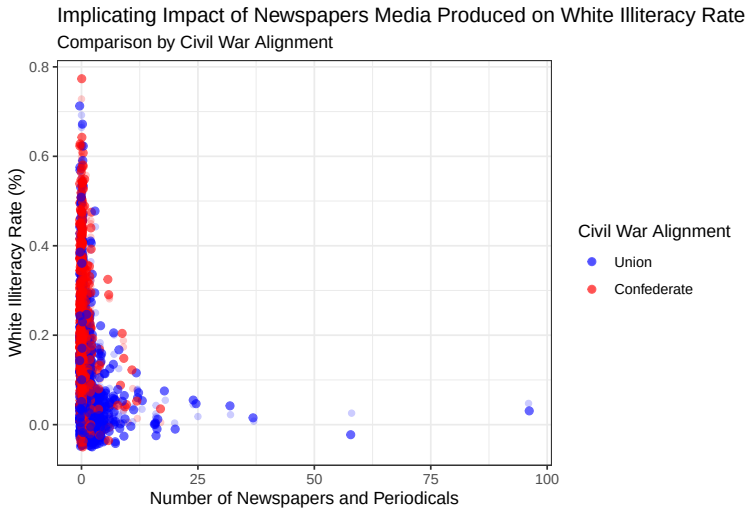
Filtering for only the columns we're interested in and creating a new list `WhiteIlliterate`

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

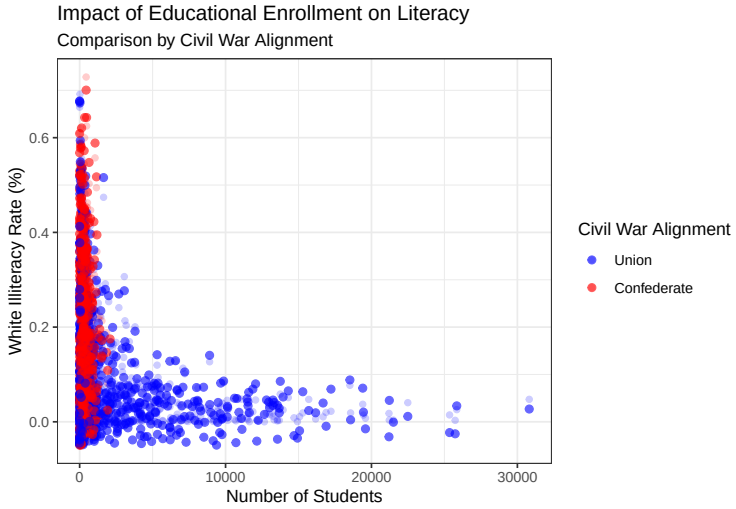
The visualization of the response variable, White Illiteracy, with Percentage of Non-White Population as potential predictor



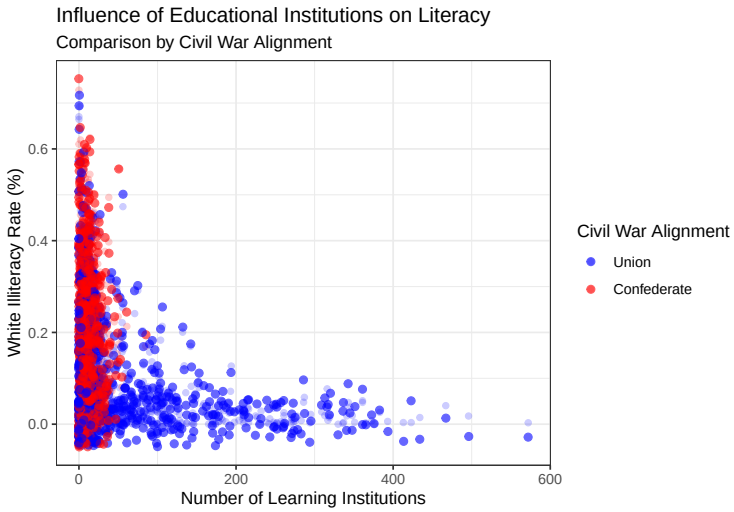
Visualizing the response variable, White Illiteracy, with Number of Newspapers and Periodicals produced as potential predictor



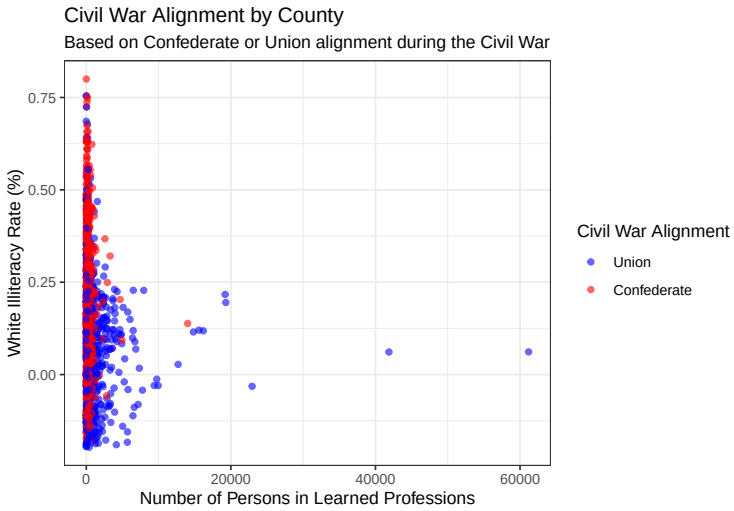
Visualizing the response variable, White Illiteracy, with Number of Students as potential predictor



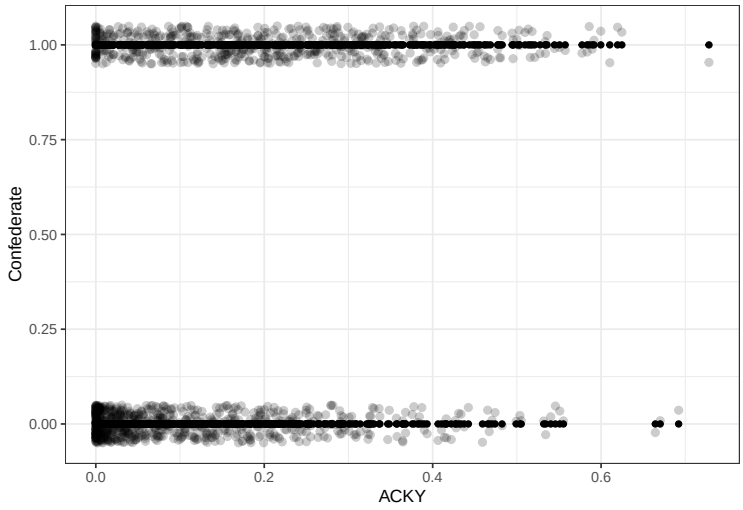
Visualizing the relationship between the Number of Learning Institutions and White Illiteracy



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] 0.1483997
```

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=-1118.7
## ACKY ~ (STATE + COUNTY + STATEICP + ACD001 + ACN001 + ACY + ACZ +
##     `Total - ACZ` + 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 +
```

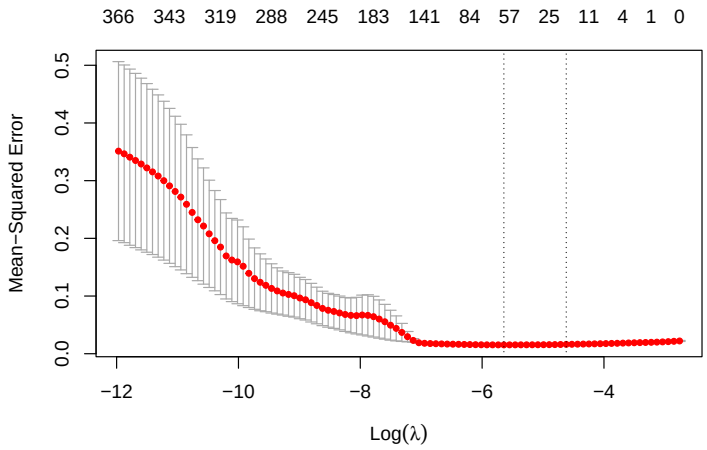
```
##     ACU001 + ACU002 + ACU003 + ACU004 + ACV001 + ACV002 + ACX001 +
##     AC4001 + AC7001 + AC8001 + AC9001 + AC9002 + AC9003 + AC9004 +
##     AC9005 + AC9006 + AC9007 + AC9008 + AC9009 + AC9010 + AC9011 +
##     AC9012 + AC9013 + AC9014 + AC9015 + AC9016 + AC9017 + AC9018 +
##     AC9019 + AC9020 + AC9021 + AC9022 + AC9023 + AC9024 + AC9025 +
##     AC9026 + AC9027 + AC9028 + AC9029 + AC9030 + ADA001 + ADA002 +
##     ADA003 + ADA004 + ADA005 + ADA006 + ADA007 + ADA008 + ADA009 +
##     ADA010 + ADA011 + ADA012 + ADA013 + ADA014 + ADA015 + ADA016 +
##     ADA017 + ADA018 + ADA019 + ADA020 + ADA021 + ADA022 + ADA023 +
##     ADA024 + ADA025 + ADA026 + ADA027 + ADA028 + ADA029 + ADA030 +
##     ADA031 + ADA032 + ADA033 + ADA034 + ADA035 + ADA036 + ADA037 +
##     ADA038 + ADA039 + ADA040 + ADA041 + 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 + 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 + 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 + AC5001 + AC5002 + AC5003 +
##     AC5004 + AC5005 + AC5006 + AC5007 + AC5008 + AC5009 + AC6001 +
##     AB2001 + AB2002 + AB2003 + AB2004 + AB3001 + AB4001 + AB4002 +
##     AB4003 + AB4004 + AB4005 + AB4006 + AB4007 + AB4008 + AB4009 +
##     AB4010 + AB4011 + AB4012 + AB4013 + AB4014 + AB4015 + AB4016 +
##     AB4017 + AB4018 + AB4019 + AB4020 + AB4021 + AB4022 + AB5001 +
##     AB5002 + AB5003 + AB5004 + AB5005 + AB6001 + AB6002 + AB6003 +
##     AB6004 + AB7001 + AB7002 + AB7003 + AB7004 + AB7005 + AB7006 +
##     AB7007 + AB7008 + AB7009 + AB7010 + AB7011 + AB7012 + AB7013 +
##     AB7014 + AB7015 + AB7016 + AB7017 + AB7018 + AB7019 + AB7020 +
##     AB7021 + AB7022 + AB8001 + AB8002 + AB8003 + AB8004 + AB8005 +
##     AB8006 + AB8007 + AB8008 + AB8009 + AB8010 + AB8011 + AB8012 +
##     AB8013 + AB8014 + AB8015 + AB8016 + AB8017 + AB8018 + AB8019 +
##     AB8020 + AB8021 + AB8022 + AB9001 + ACA001 + ACA002 + ACA003 +
##     ACB001 + ACC001 + ACZYZ + ACH + ACI + ACL + ACM + AC1 + ACE +
##     AC3 + AC9M + AC9C + AC9MF + ACS + ADAF + ADAMF + ADAM + ADCM +
##     ADCC + ADCMFF + ADCMFM + ADCMFT + ADCMFTH + ADCMFL + ADCMFSCL +
##     ADCMFC + ADCMFGE + ADCMFFP + ADCMFPP + ADCMFO + ADEM + ADEC +
##     ADEMF + AC5 + AB2 + AB5 + AB8 + ACSW + ACSNW + ACSP + ACSWP +
##     Confederate) - STATE - STATEICP - COUNTY

##             Estimate   Std. Error   t value     Pr(>|t|)
## AB2002  4.245243e-06 1.315045e-06  3.228211 1.312053e-03
## AB4017 -5.002397e-05 9.664333e-06 -5.176143 3.073287e-07
```

Forward Stepwise Selection using OLS led to just `AB2002` and `AB4017` as predictor variables. We will use these variables to build a multiple linear regression model.

```
## [1] 0.1361309
```

And our Forward Stepwise selection method using AIC as criteria gave us a model using `AB2002` and `AB4017` as predictor variables, with a testing RMSE of 0.1361309**.**

Ridge and Lasso Regression (plus OLS glm)




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

```
## [1] 0.003542939
```

```
## [1] 0.1248152
```

```
## [1] 0.7520277
```

```
## [1] 0.1524679
```

Our glm model using Ordinary Least Squares model yielded RMSE of 0.7520277. Our model using Ridge Regression model yielded RMSE of 0.1524679. Our best model was using Lasso Regression model with testing RMSE of 0.1248152.

building full lasso model using all data

```
## [1] 0.04713318
```

```
##        ACZ014        ACZ015        AC3001        ACE001        ACE002
## -2.095020e-05 -2.306236e-05  2.218967e-03  5.790964e-04  9.238531e-05
##        ACF002        ACF003        ACG001        ACH001        ACH003
##  6.610901e-05  2.596812e-03  4.088632e-04 -5.048432e-03 -2.824031e-05
##        ACM002        ACX001        AC9021        AC9022        AC9023
```

```
## -5.988777e-03 -1.527420e-02 -3.188013e-03 -1.015925e-03  5.545981e-04
##        AC9026        AC9027        AC9030        ADA002        ADA005
##  4.955895e-04  2.953062e-04 -1.401204e-04  1.267194e-07  1.222929e-08
##        ADA008        ADA015        ADB003        ADB004        ADB018
##  8.101391e-08 -1.029596e-07  9.827415e-09  7.460073e-07  2.797926e-09
##        ADB025        ADB028        ADC006        ADC009        ADE007
##  2.057057e-06 -6.529832e-05  6.030814e-08 -2.685067e-09  1.040954e-04
##        ADE011        ADE014        AB2004        AB3001        AB4001
##  3.986762e-06  7.284393e-06  3.549597e-07 -1.996919e-07 -4.585835e-08
##        AB4009        AB4011        AB4017        AB5003        AB5005
##  2.007877e-05  2.171011e-08 -3.775408e-11  8.085304e-09  1.821483e-07
##        AB7003        AB7005        AB7008        AB7011        AB7013
##  1.343076e-01 -9.169049e-02 -1.039481e-01 -1.770470e-02 -1.326899e-03
##        AB7020        AB7022        AB8005        AB8010        AB8011
##  3.781912e-04 -6.357559e-02 -6.122177e-07 -1.106676e-07  1.868205e-07
##        AB8017        AB9001           ACH       ADCMFGE          ACSP
## -3.737099e-11 -1.481212e-08 -3.874908e-05 -1.210990e-08 -4.548443e-02
##         ACSWP
##  4.713318e-02
```

```
##  [1] "ACZ014" "ACZ015" "AC3001" "ACE001" "ACE002" "ACF002" "ACF003" "ACG001"
##  [9] "ACH001" "ACH003" "ACM002" "ACX001" "AC9021" "AC9022"
```

For ACKY as the response variable and only using the numerical data in the predictive model, but using all data, we get the following predictor variables with non-zero coefficients: “ACZ015”, “ACM002”, “ACX001”, “AC9022”, “AC9023”, “AC9027”, “ADA015”, “ADB028”, “ADC037”, “AB4009”, “AB7008”, “AB7013”, “AB7020”, and “AB9001”.

```
##     ACZ015             ACM002             ACX001             AC9022
##  Length:1           Length:1           Length:1           Length:1
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     AC9023             AC9027             ADA015             ADB028
##  Length:1           Length:1           Length:1           Length:1
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     ADC037             AB4009             AB7008             AB7013
##  Length:1           Length:1           Length:1           Length:1
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     AB7020             AB9001
##  Length:1           Length:1
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

```
##                                                 ACZ015
##                    "Slave >> Male >> 24 to 35 years of age"
##                                                 ACM002
##                                  "Weekly newspapers"
##                                                 ACX001
##                                          "0=NO 1=YES"
##                                                 AC9022
##                     "Manufacturing: Mills: Flouring mills"
```

```
##                                                      AC9023
##                            "Manufacturing: Mills: Grist mills"
##                                                      AC9027
## "Manufacturing: Distilled and Fermented Liquors: Distilleries"
##                                                      ADA015
##    "Manufacturing Products: Earthenware manufactured articles"
##                                                      ADB028
##                       "Printing and Binding: Weekly newspapers"
##                                                      ADC037
##                           "Manufacturing Establishments: Mills"
##                                                      AB4009
##                              "Various crops: Beeswax (pounds)"
##                                                      AB7008
##                                 "Various crops: Hops (pounds)"
##                                                      AB7013
##                                   "Various crops: Hemp (tons)"
##                                                      AB7020
##               "Cotton, silk, sugar, etc.: Cane sugar (pounds)"
##                                                      AB9001
##                                  "Estimated crop output value"
```

For ACKY as the response variable, Lasso gave us a model with the following predictor variables:

- ACZ015 - Slave » Male » 24 to 35 years of age
- ACM002 - Weekly newspapers
- ACX001 - 0=NO 1=YES (Navigable Waterway)
- AC9022 - Total Number of Establishments: Manufacturing: Mills: Flouring mills
- AC9023 - Total Number of Establishments: Manufacturing: Mills: Grist mills
- AC9027 - Total Number of Establishments: Manufacturing: Distilled and Fermented Liquors: Distilleries
- ADA015 - Total Value of Production of Establishments: Manufacturing Products: Earthenware manufactured articles
- ADB028 - Total Produced Number of Printing and Binding: Weekly newspapers
- ADC037 - Capital Invested in Manufacturing Establishments: Mills
- AB4009 - Various crops: Beeswax (pounds)
- AB7008 - Various crops: Hops (pounds)
- AB7013 - Various crops: Hemp (tons)
- AB7020 - Cotton, silk, sugar, etc.: Cane sugar (pounds)
- AB9001 - Estimated total crop output value

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] 0.1184195
```

The full lasso model had an RMSE of 0.1184195 **using the best lambda value of** 0.003542939**. This is an improvement over the testing lasso RMSE of** 0.1248152**.**

best subset

For Paper 1 stuff

Histograms of predictor variables distributions