Presentation

Minsu Kim, Ariane Stark

1. Introduction
2. Simple Linear Regression
3. Subset Selection
4. Shrinkage Methods
5. Tree Based Methods
6. Classification
7. Conclusion

Section 1

Introduction

Background

Study Objectives

-
- Fit a classification model to predict whether a given county has an instance rate higher than the median of the instance rates.

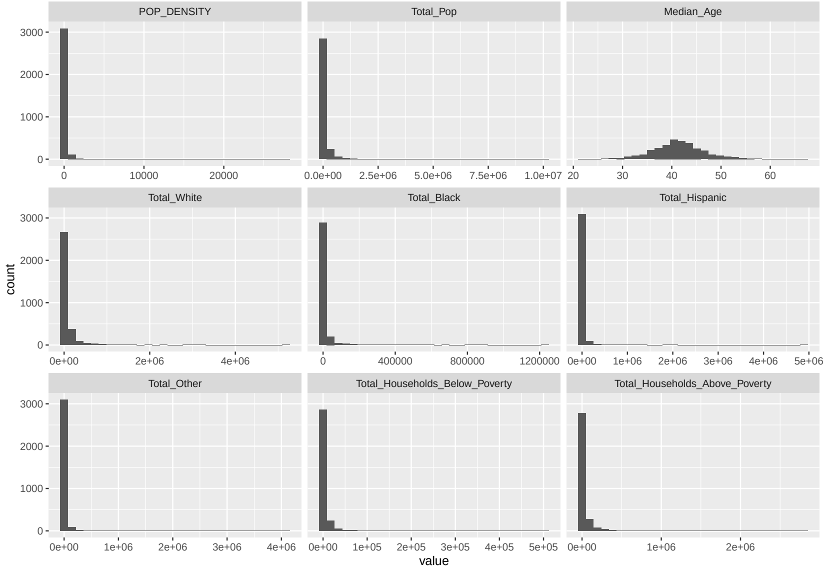
Data Set Overview

The data set contains cumulative COVID-19 case counts as of April 1, 2021 for each of the 3,220 counties in the United States. Also we have racial and ethnicity demographics, poverty demographics, population demographics, and age demographics from government census data. Note Rio Arriba County, New Mexico is missing poverty demographics and has thus been excluded from analysis. leaving our data set with 9 predictors and 3219 observations

Our Variables

- Cumulative COVID-19 Cases
- Total Population
- Population Density
- Median Age
- Total Number of White Residents
- Total Number of Black Residents
- Total Number of Other (Not strictly White or Black) Residents
- Total Number of Hispanic Residents
- Total Number of Households Above the Poverty Line
- Total Number of Households Below the Poverty Line

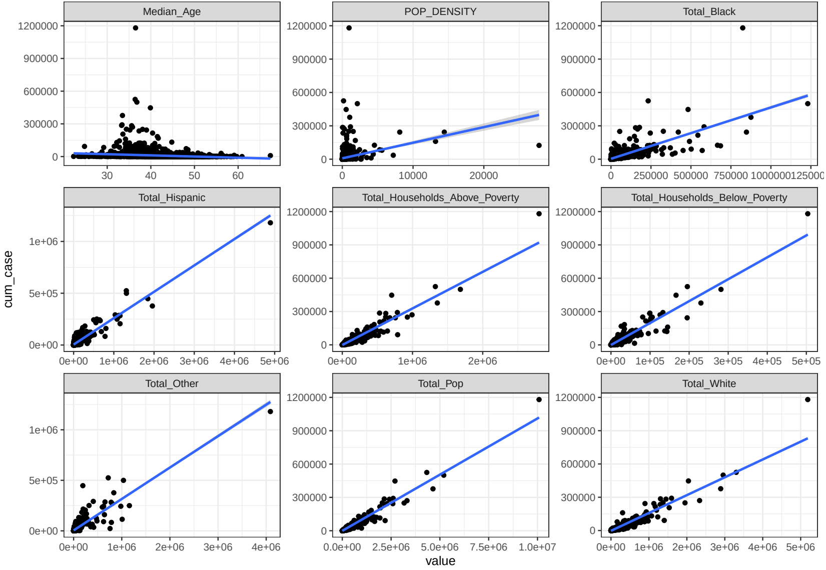
Our Variables: Plots



Section 2

Simple Linear Regression

Simple Linear Regression



Slide Name

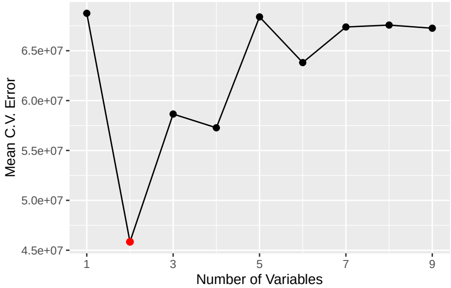
slide text

Section 3

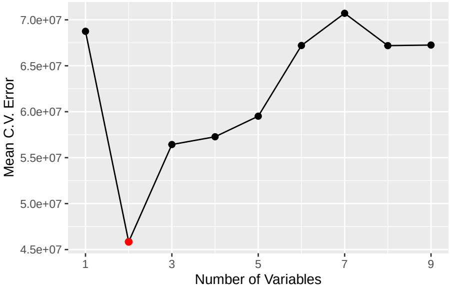
Subset Selection

Cross Validation of Best Subset, Forward and Backward Stepwise

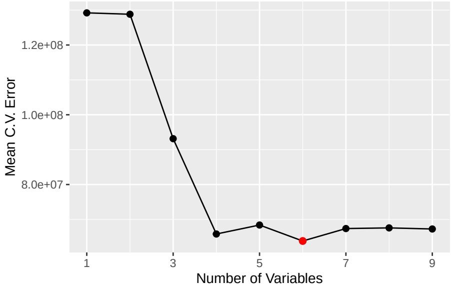
10-fold cross validation of data was done to find the optimal number of parameters for the model selected using each method. This is done to balance the model fitting the data well and the model having decent predictive accuracy.



Best Subset Selection



Forward Stepwise Selection



Backward Stepwise Selection

Best Subset Overall

Recall 10 fold cross validation of best subset selection selected the model with 2 parameters as having the lowest average cross validation predictive error.

Parameters	R^2	AdjR^2	CP	BIC
1	0.9452	0.9452	2488.4481	-9330.249
2	0.9639	0.9638	546.0689	-10664.172
3	0.9662	0.9662	306.2665	-10870.117
4	0.9674	0.9674	178.4008	-10983.143
5	0.9683	0.9683	89.4799	-11062.648
6	0.9691	0.9691	5.6465	-11139.497
7	0.9691	0.9691	6.1421	-11132.929
8	0.9691	0.9691	8.0746	-11124.920
9	0.9692	0.9691	10.0000	-11116.918

Forward Stepwise Overall

Recall 10 fold cross validation of forward stepwise selection selected the model with 2 parameters as having the lowest average cross validation predictive error.

Parameters	R^2	AdjR^2	CP	BIC
1	0.9452	0.9452	2488.4481	-9330.249
2	0.9639	0.9638	546.0689	-10664.172
3	0.9662	0.9662	306.2665	-10870.117
4	0.9674	0.9674	178.4008	-10983.143
5	0.9675	0.9674	175.8868	-10979.359
6	0.9691	0.9691	5.6465	-11139.497
7	0.9691	0.9691	6.1421	-11132.929
8	0.9691	0.9691	8.0746	-11124.920
9	0.9692	0.9691	10.0000	-11116.918

Backward Stepwise Overall

Recall 10 fold cross validation of backward stepwise selection selected the model with 6 parameters as having the lowest average cross validation predictive error.

Parameters	R^2	AdjR^2	CP	BIC
1	0.9025	0.9025	6923.4012	-7478.489
2	0.9559	0.9559	1375.4685	-10022.384
3	0.9648	0.9647	452.3814	-10739.096
4	0.9665	0.9665	270.9103	-10896.412
5	0.9683	0.9683	89.4799	-11062.648
6	0.9691	0.9691	5.6465	-11139.497
7	0.9691	0.9691	6.1421	-11132.929
8	0.9691	0.9691	8.0746	-11124.920
9	0.9692	0.9691	10.0000	-11116.918

Model Summary

Best Subset: The model has parameters: Total Population, and Total Hispanic.

Cumulative Cases = 315.11931521 + 0.07300349(Total Population) + 0.08304726(Total Hispanic)

Forward Stepwise: The model has parameters: Total Population, Total White, and Total Hispanic.

Cumulative Cases = 315.11931521 + 0.07300349(Total Population) + 0.08304726(Total Hispanic)

Backward Stepwise: The model has parameters: Total Population, Total White, Total Black, Total Hispanic, Total Other, and Total Number of Households Above Poverty.

Cumulative Cases = -338.60412493 -0.38868473(Total Population) +0.58210948(Total White) +0.55858828(Total Black) +0.05514001(Total Hispanic) +0.51284419(Total Other) -0.29137162(Total Number of Households Above Poverty)

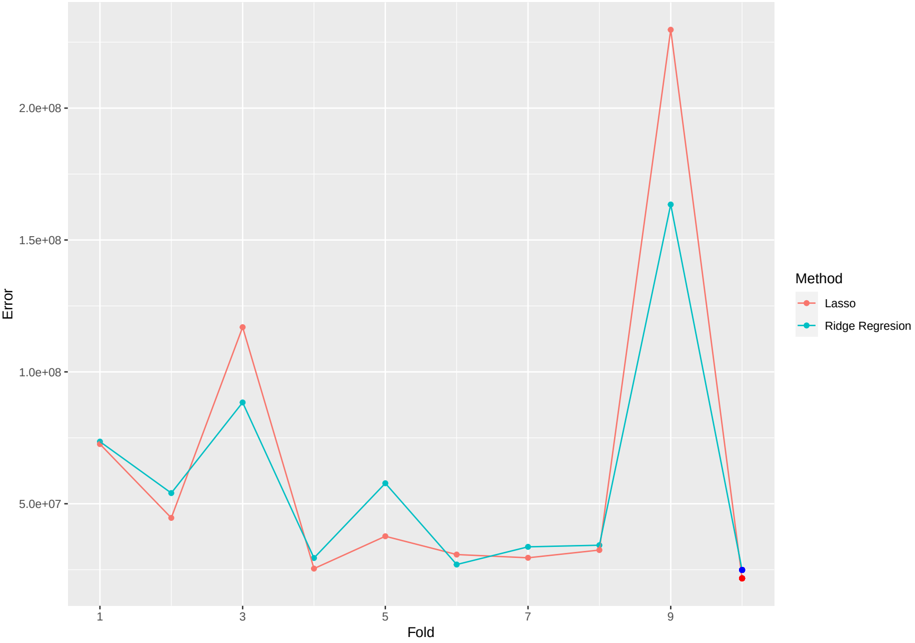
Section 4

Shrinkage Methods

Ridge Regression and Lasso

Recall that Ridge Regression performs better when the response is a function of many predictors of roughly equal size and Lasso performs better when the response is a function of a small subset of predictors.

10-fold Cross Validation was performed to find the optimal tuning parameter λ for each method.


Error
2.0e+08
1.5e+08
1.0e+08
5.0e+07
1
3
5
7
9
Fold
Method
Lasso
Ridge Regresion

Ridge With Lambda that Produces Smallest C.V. Error

The smallest Cross Validation Error for Ridge Regression is 24901393 with and average of 58650401.

The model selected is:

Cumulative Cases = 2659.26735391 -1.46826014(Population Density) +0.01675252(Total Population) -60.01296350(Median Age) +0.03389108(Total White) +0.04046944(Total Black) +0.06843143 (Total Hispanic) +0.03043294 (Total Other) +0.29137162(Total Number of Households Above Poverty) +0.27874317(Total Number of Households Below Poverty)

Lasso With Lambda that Produces Smallest C.V. Error

The smallest Cross Validation Error for Lasso is 21711998 with and average of 64144413.

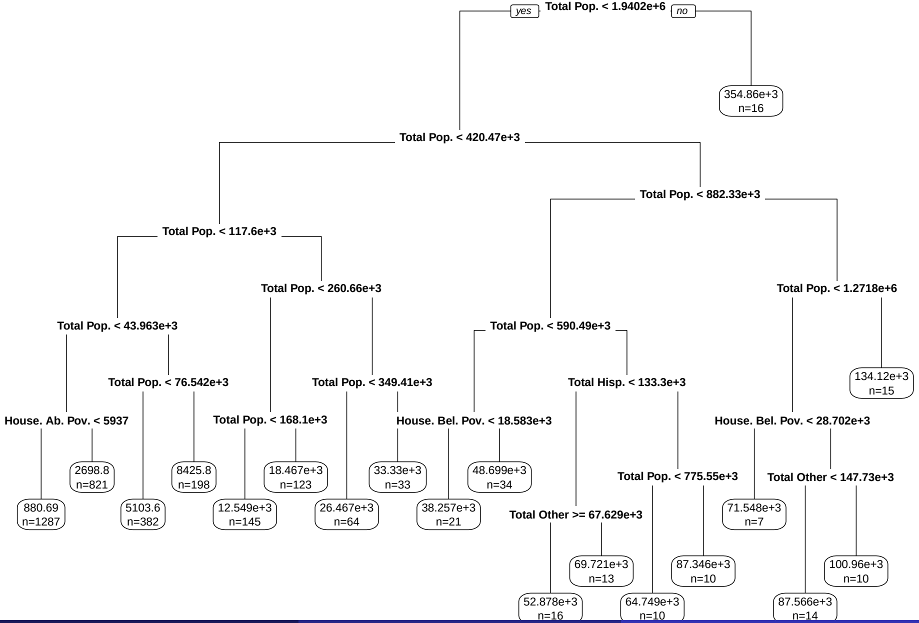
The model selected is:

Cumulative Cases $=$ 286.95400883 $+$0.04444873(Total Population) $+$0.02943025(Total White) $+$0.01685243(Total Black) $+$0.08547591(Total Hispanic) $+$0.09854572(Total Number of Households Below Poverty)

Section 5

Tree Based Methods

Regression Trees



Section 6

Classification

Slide Name

slide text

Section 7

Conclusion

Slide Name

slide text