


Source: Johns Hopkins University

ELECTION AND DEMOGRAPHICS EFFECT ON COVID-19 CASES

A STATISTICAL LEARNING FOR
MACHINE LEARNING ANALYSIS

By Michael Sanky and Ely Feintuch

INTRODUCTION

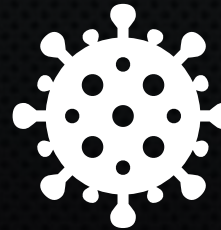


LASSO, RIDGE, ELASTIC
NET, AND RANDOM
FORESTS WILL BE USED FOR
MACHINE LEARNING
REGRESSION ANALYSIS

This analysis is based on a dataset of the effect of counties' demographics, economic statistics, and presidential election results on covid-19 cases. It can be found at:

<https://www.Kaggle.Com/etsc9287/2020-general-election-polls>

We took a sampling of 1,468 counties and utilize 39 numerical predictors and 1 categorical predictor. We predict the log of cases per county. Some predictors include percentage of residents who commute via public transport, percentage voted for Hillary Clinton, and average income.

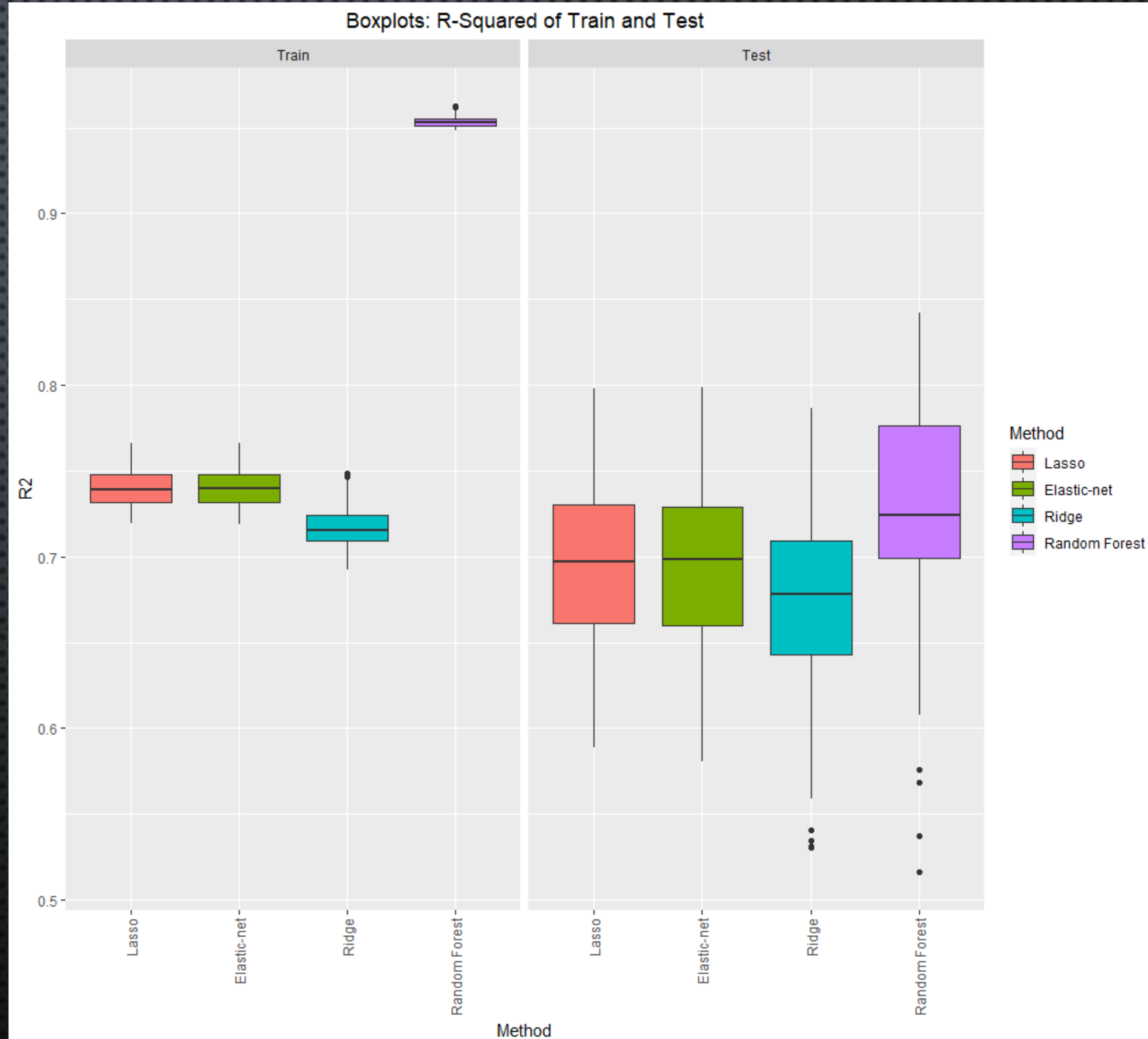


$N = 1,468$

$P = 40$

Boxplots of Train and Test R^2

- Data is split 80-20 into Training and Test sets, repeated 100 times.
- Lambdas are tuned using 10-fold CV.
- Random Forest has the highest average R-squared.



Performance times:

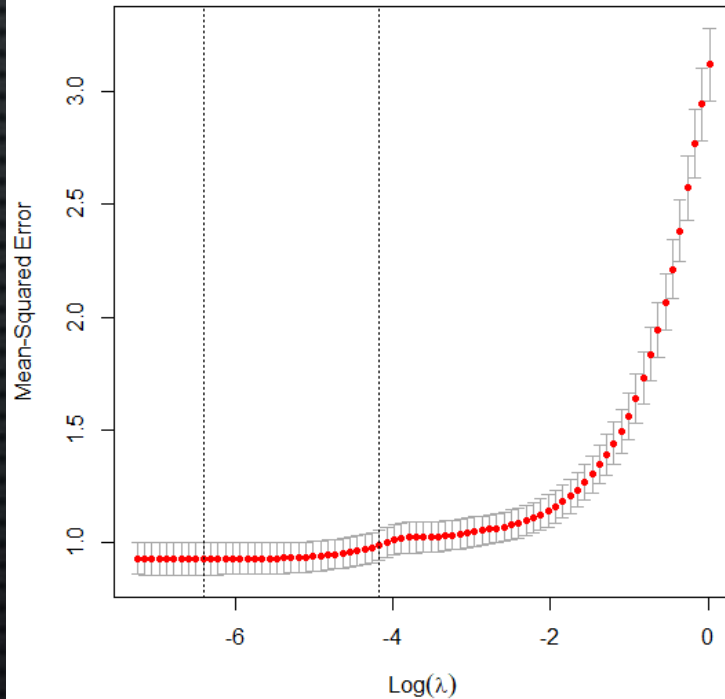
- Lasso: 0.31 seconds
- Elastic Net: 0.29 seconds
- Ridge: 0.20 seconds

10-FOLD CROSS VALIDATION CURVES

For a single sample, 10-fold cross validation curves are given.

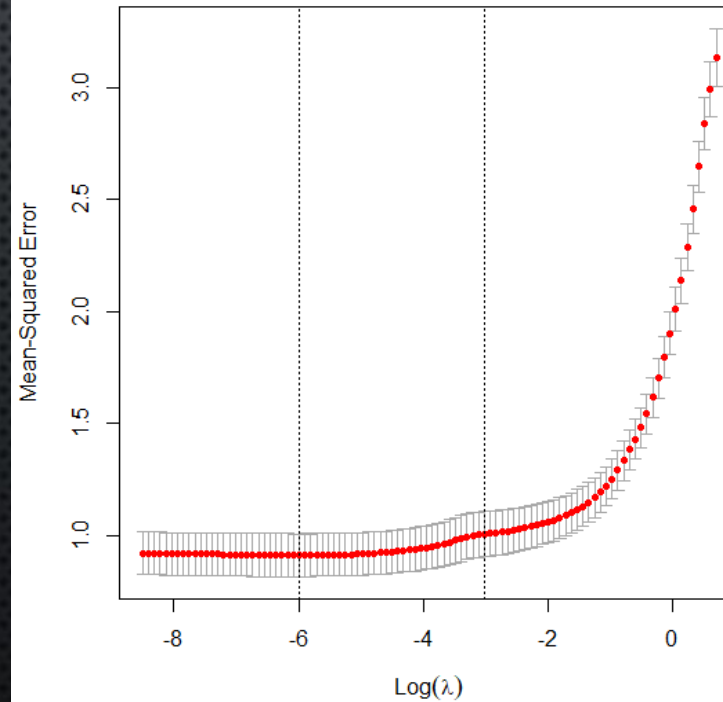
10-fold CV curve for Lasso

38 35 34 32 32 27 24 20 19 16 11 9 8 4



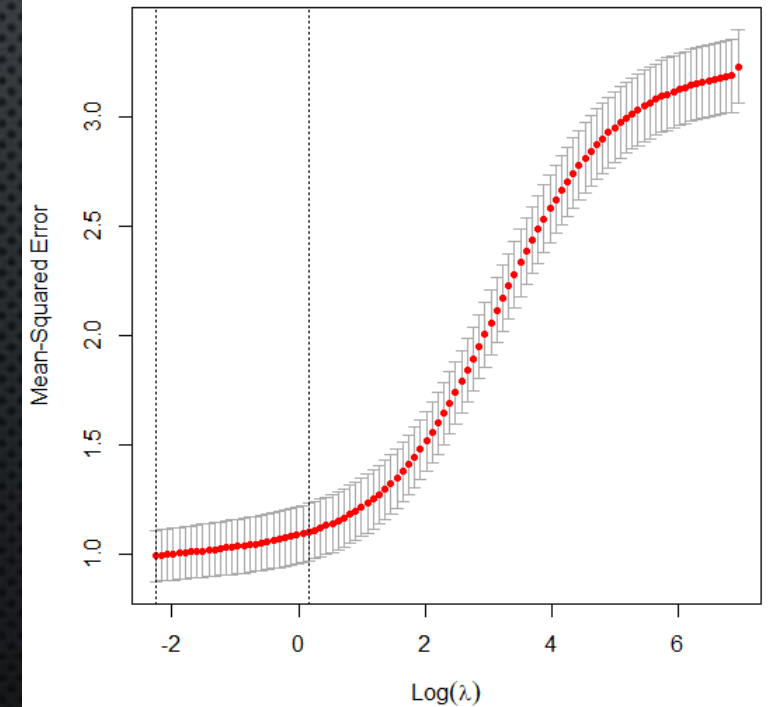
10-fold CV curve for Elastic-net

39 39 39 35 34 33 29 28 22 20 14 12 8



10-fold CV curve for Ridge

40 40 40 40 40 40 40 40 40 40 40 40



Boxplots of Train and Test Residuals of a Single Sample



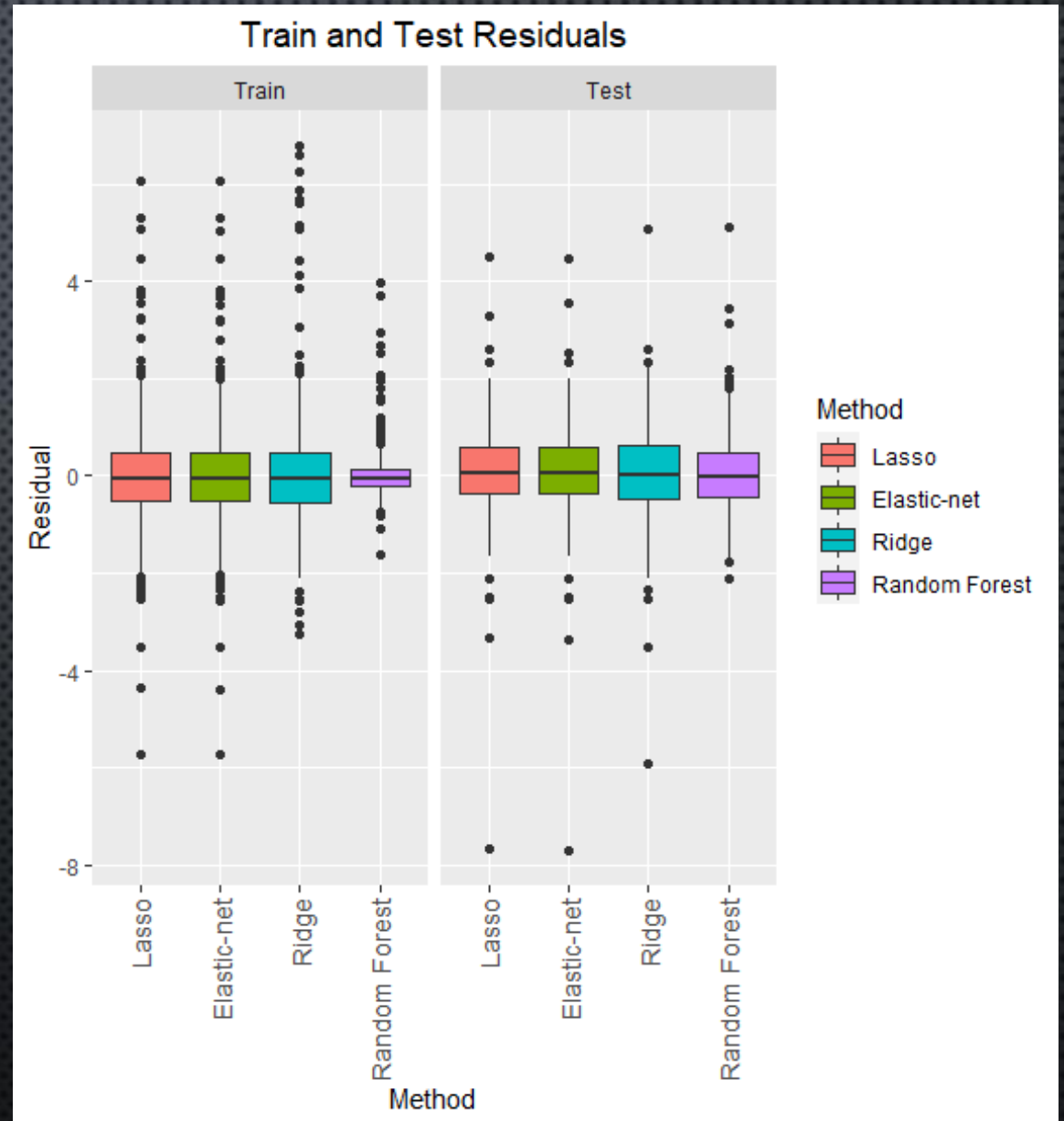
All 4 methods have residuals close to zero.



Lasso and Elastic-net are virtually indistinguishable.



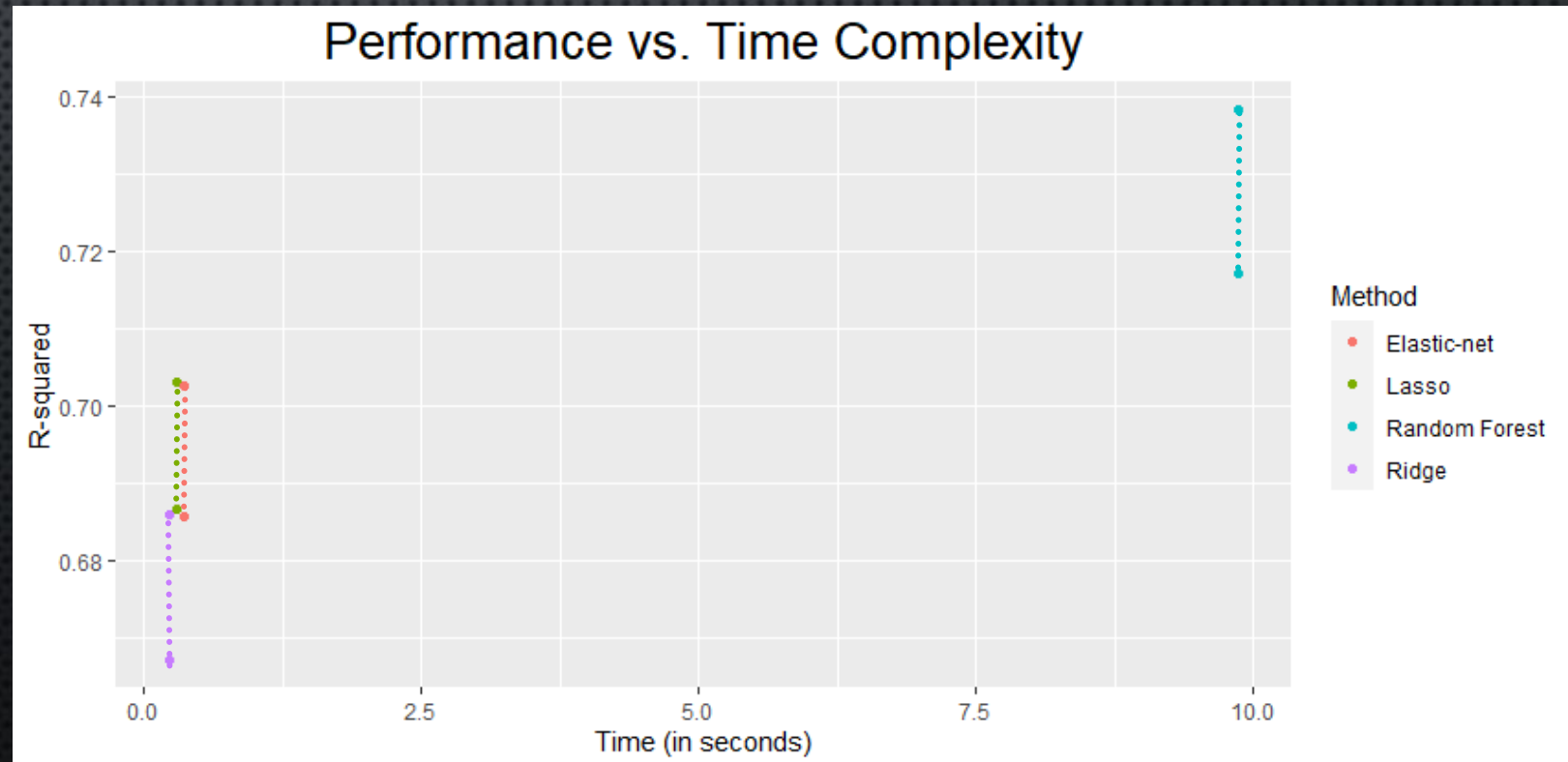
The test residuals from Random Forests are smaller than from other methods.



R² vs. Time

- Random Forest performed only marginally better, with a cost of being 33x slower.
- Ridge, Lasso, and Elastic-net have comparable performances and length of computation.
- Of the three, Ridge is quickest and has the lowest R².

Method	90% Test R ² Interval	Elapsed Time
Lasso	{0.687, 0.703}	0.30 sec
Elastic-net	{0.686, 0.702}	0.36 sec
Ridge	{0.667, 0.686}	0.23 sec
Random Forest	{0.717, 0.738}	9.86 sec



PARAMETER IMPORTANCE



Elastic-net and Lasso select similarly, while Ridge selects more parameters.

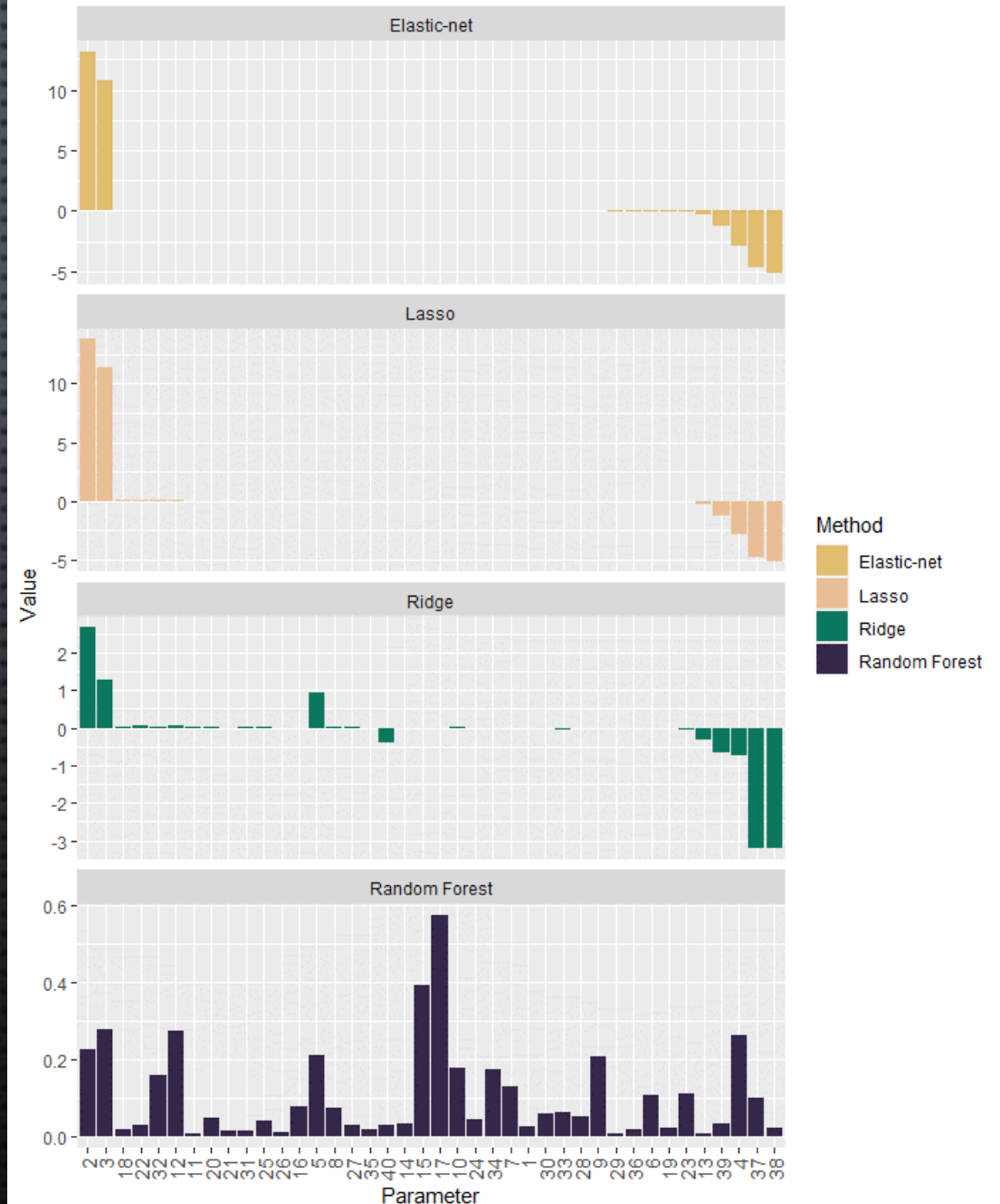


For Lasso/Elastic-net/Ridge, the parameters with higher magnitude coefficients mostly regard voting.



Along with voting patterns, Random Forest considers economic factors as most important.

Parameter Importance By Method



CLOSING REMARKS

Political and demographic factors have significant association with COVID-19 case count.

Lasso and Elastic-net act similarly on this data, having similar CV curves, time complexities, coefficients and residuals.

Random Forest values variables differently, and what it lacks in time complexity it makes up in its high R-squared and residuals closer to zero.

The full R code for this project can be accessed at:

https://github.com/michaelsanky/Public-Codes/blob/main/STA%209890_Sanky_Feintuch.R