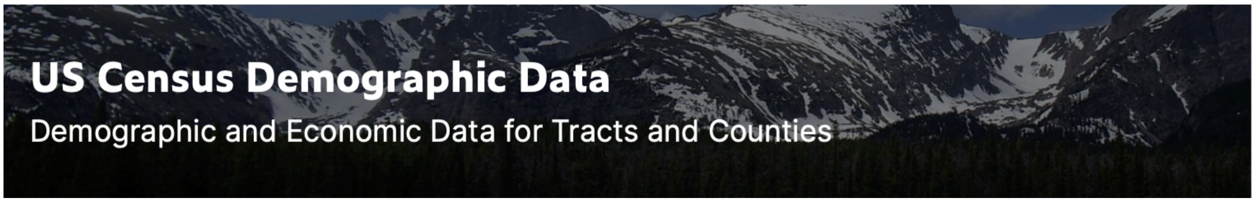
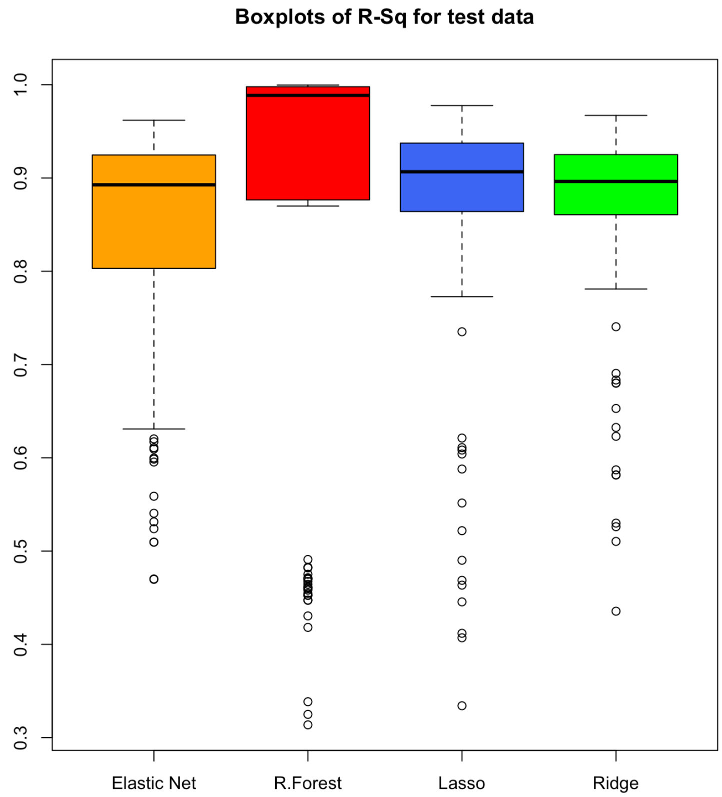
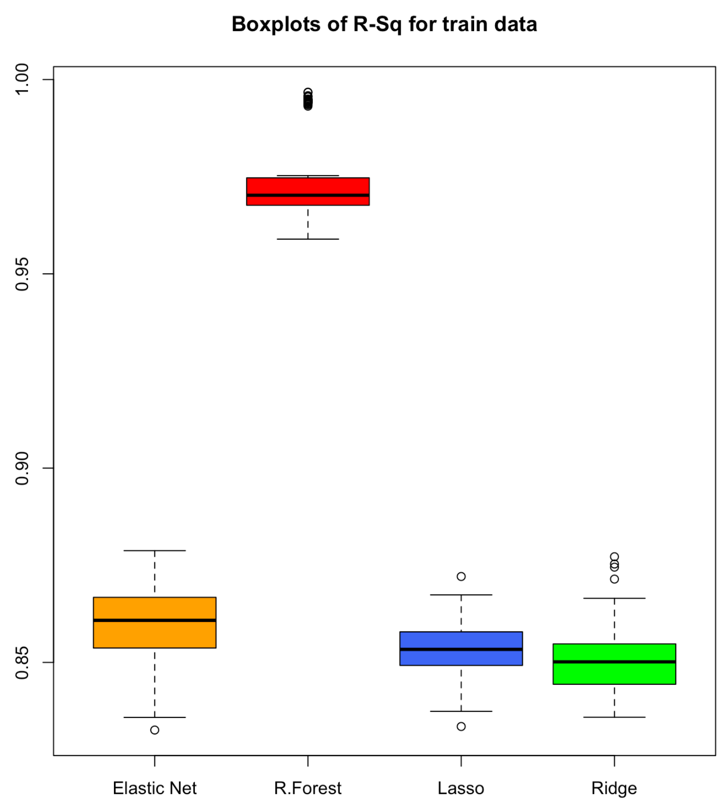
Krist Kikina
STA 9890
Prof. Kamiar Rahnama Rad



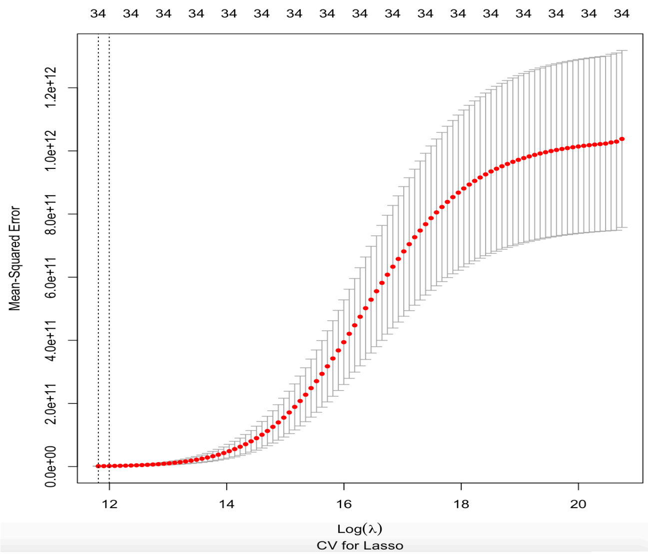

```
Observations: 3,220
Variables: 37
$ CountyId         <dbl> 1001, 1003, 1005, 1007, 1009, 1011, 1013, 1015, 1017, 1019, 1021, 1023, 1025,…
$ State            <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Alabama", "Alabama", "Alabama", …
$ County           <chr> "Autauga County", "Baldwin County", "Barbour County", "Bibb County", "Blount …
$ TotalPop         <dbl> 55036, 203360, 26201, 22580, 57667, 10478, 20126, 115527, 33895, 25855, 43805…
$ Men              <dbl> 26899, 99527, 13976, 12251, 28490, 5616, 9416, 55593, 16320, 12862, 21554, 62…
$ Women            <dbl> 28137, 103833, 12225, 10329, 29177, 4862, 10710, 59934, 17575, 12993, 22251, …
$ Hispanic         <dbl> 2.7, 4.4, 4.2, 2.4, 9.0, 0.3, 0.3, 3.6, 2.2, 1.6, 7.7, 0.5, 0.2, 3.1, 2.4, 6.…
$ White            <dbl> 75.4, 83.1, 45.7, 74.6, 87.4, 21.6, 52.2, 72.7, 56.2, 91.8, 80.4, 56.3, 53.0,…
$ Black            <dbl> 18.9, 9.5, 47.8, 22.0, 1.5, 75.6, 44.7, 20.4, 39.3, 5.0, 9.5, 42.1, 45.7, 14.…
$ Native           <dbl> 0.3, 0.8, 0.2, 0.4, 0.3, 1.0, 0.1, 0.2, 0.3, 0.5, 0.4, 0.0, 0.1, 0.9, 0.3, 1.…
$ Asian            <dbl> 0.9, 0.7, 0.6, 0.0, 0.1, 0.7, 1.1, 1.0, 1.0, 0.1, 0.4, 0.1, 0.5, 0.0, 0.5, 1.…
$ Pacific          <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.1, 0.0, 0.0, 0.…
$ VotingAgeCitizen <dbl> 41016, 155376, 20269, 17662, 42513, 8212, 15459, 88383, 26259, 20620, 31776, …
$ Income           <dbl> 55317, 52562, 33368, 43404, 47412, 29655, 36326, 43686, 37342, 40041, 43501, …
$ IncomeErr        <dbl> 2838, 1348, 2551, 3431, 2630, 5376, 2701, 1491, 2011, 2316, 2877, 2797, 2336,…
$ IncomePerCap     <dbl> 27824, 29364, 17561, 20911, 22021, 20856, 19004, 23638, 22002, 23010, 23368, …
$ IncomePerCapErr  <dbl> 2024, 735, 798, 1889, 850, 2355, 943, 793, 1205, 1354, 1925, 1307, 1203, 1553…
$ Poverty          <dbl> 13.7, 11.8, 27.2, 15.2, 15.6, 28.5, 24.4, 18.6, 18.8, 16.1, 19.4, 22.3, 25.3,…
$ ChildPoverty     <dbl> 20.1, 16.1, 44.9, 26.6, 25.4, 50.4, 34.8, 26.6, 29.1, 20.0, 27.8, 32.8, 30.7,…
$ Professional     <dbl> 35.3, 35.7, 25.0, 24.4, 28.5, 19.7, 26.9, 29.0, 24.3, 28.8, 25.3, 23.6, 21.6,…
$ Service          <dbl> 18.0, 18.2, 16.8, 17.6, 12.9, 17.1, 17.3, 17.5, 13.5, 14.8, 14.5, 15.4, 14.3,…
$ Office           <dbl> 23.2, 25.6, 22.6, 19.7, 23.3, 18.6, 18.5, 23.7, 23.0, 18.1, 23.7, 22.0, 24.8,…
$ Construction     <dbl> 8.1, 9.7, 11.5, 15.9, 15.8, 14.0, 11.6, 10.4, 11.6, 11.9, 15.5, 17.1, 13.7, 1…
$ Production       <dbl> 15.4, 10.8, 24.1, 22.4, 19.5, 30.6, 25.7, 19.4, 27.6, 26.5, 21.0, 21.9, 25.6,…
$ Drive            <dbl> 86.0, 84.7, 83.4, 86.4, 86.8, 73.1, 83.6, 85.0, 87.1, 85.0, 83.2, 81.8, 83.7,…
$ Carpool          <dbl> 9.6, 7.6, 11.1, 9.5, 10.2, 15.7, 12.6, 9.2, 9.7, 12.1, 12.6, 13.7, 11.9, 6.0,…
$ Transit          <dbl> 0.1, 0.1, 0.3, 0.7, 0.1, 0.3, 0.0, 0.2, 0.2, 0.4, 0.1, 0.0, 0.2, 0.0, 0.0, 0.…
$ Walk             <dbl> 0.6, 0.8, 2.2, 0.3, 0.4, 6.2, 0.9, 1.3, 0.6, 0.3, 0.6, 1.7, 0.7, 2.8, 0.9, 1.…
$ OtherTransp      <dbl> 1.3, 1.1, 1.7, 1.7, 0.4, 1.7, 0.9, 1.1, 0.5, 0.3, 1.8, 1.2, 2.7, 0.6, 0.1, 1.…
$ WorkAtHome       <dbl> 2.5, 5.6, 1.3, 1.5, 2.1, 3.0, 2.0, 3.2, 2.0, 2.0, 1.7, 1.6, 0.9, 3.0, 2.7, 2.…
$ MeanCommute      <dbl> 25.8, 27.0, 23.4, 30.0, 35.0, 29.8, 23.2, 24.8, 23.6, 26.5, 32.5, 32.7, 23.9,…
$ Employed         <dbl> 24112, 89527, 8878, 8171, 21380, 4290, 7727, 47392, 14527, 9879, 17675, 4301,…
$ PrivateWork      <dbl> 74.1, 80.7, 74.1, 76.0, 83.9, 81.4, 79.1, 74.9, 84.5, 74.8, 81.1, 79.9, 83.1,…
$ PublicWork       <dbl> 20.2, 12.9, 19.1, 17.4, 11.9, 13.6, 15.3, 19.9, 11.8, 17.1, 14.0, 14.8, 11.8,…
$ SelfEmployed     <dbl> 5.6, 6.3, 6.5, 6.3, 4.0, 5.0, 5.3, 5.1, 3.7, 8.1, 4.5, 4.9, 5.1, 7.7, 8.2, 5.…
$ FamilyWork       <dbl> 0.1, 0.1, 0.3, 0.3, 0.1, 0.0, 0.3, 0.1, 0.0, 0.0, 0.4, 0.4, 0.0, 0.0, 0.0, 0.…
$ Unemployment     <dbl> 5.2, 5.5, 12.4, 8.2, 4.9, 12.1, 7.6, 10.1, 6.4, 5.3, 6.7, 9.8, 15.2, 6.4, 7.8…
```

Side by Side Boxplots of R^2train and R^2test



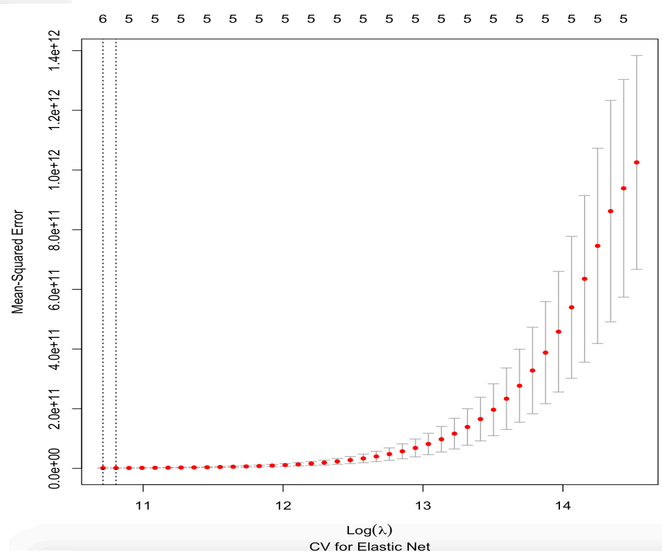


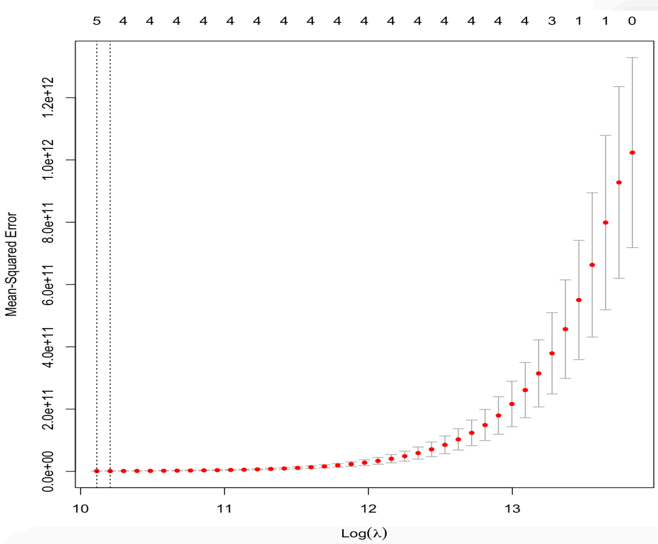
10-fold CV curves for lasso, elastic-net α = 0.5, ridge.



Lasso

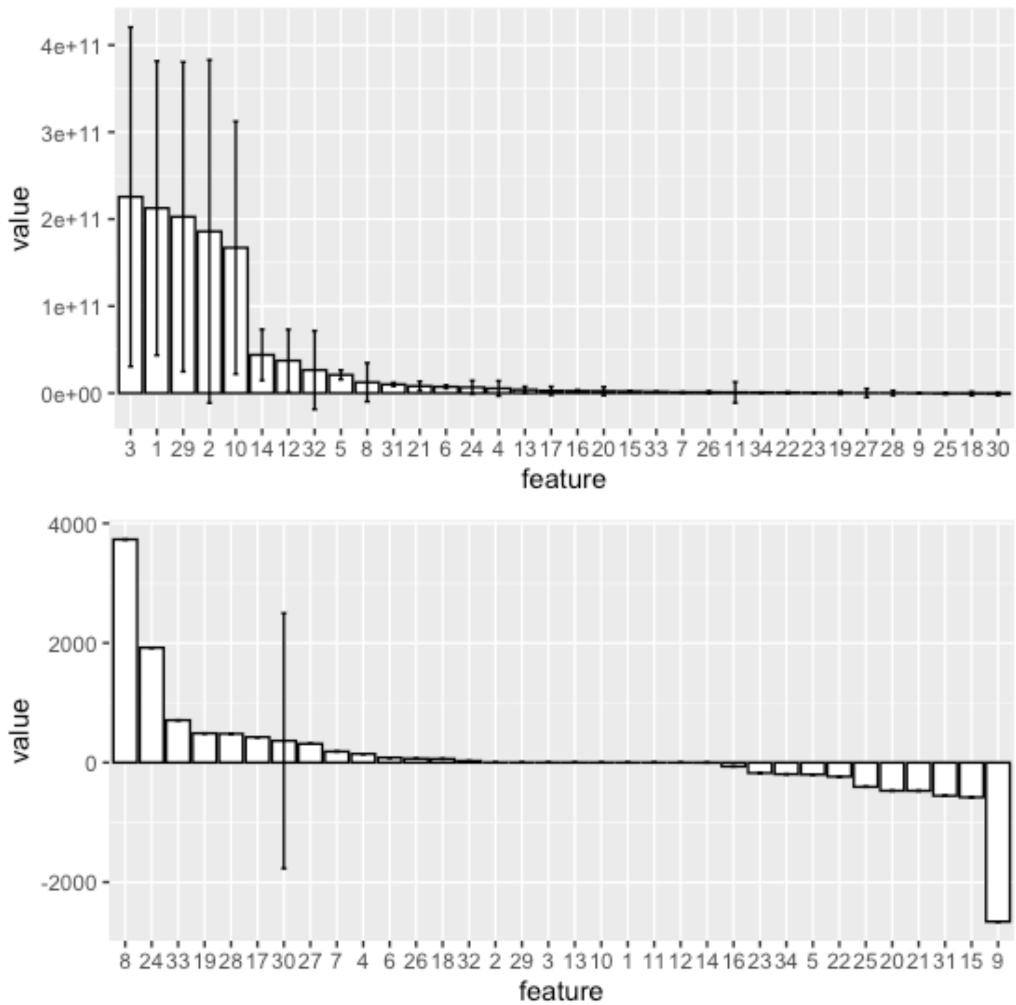
Elastic-Net





Ridge

Bar-plots (with bootstrapped error bars) of the estimated coefficients, and the importance of the parameters. If you have something interesting to say about coefficients that are (or are not important) say it.



➔ order of features that have a strong impact and response.

Summary:

Picking winning model using cross -validation (comparing elastic net, ridge, and lasso)

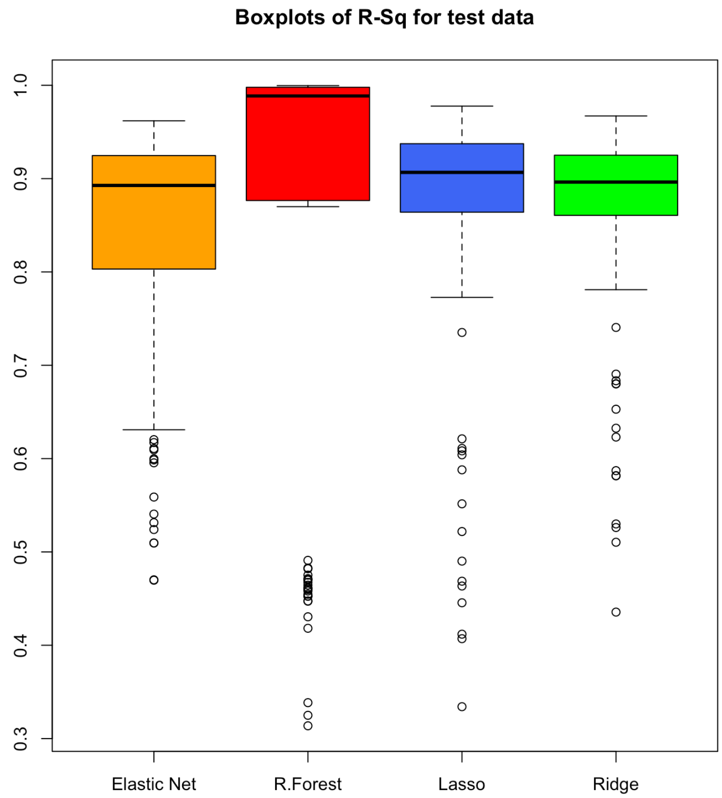
```
> cv.fit

Call:  cv.glmnet(x = X, y = y, nfolds = 10, alpha = a)

Measure: Mean-Squared Error

    Lambda  Measure         SE Nonzero
min  44927 1.127e+09 295970910       6
1se  49307 1.320e+09 360236475       6
> cv.fit.la

Call:  cv.glmnet(x = X, y = y, nfolds = 10, alpha = b)

Measure: Mean-Squared Error

    Lambda  Measure         SE Nonzero
min 134664 1.215e+09 549702326      34
1se 162203 1.693e+09 779582657      34
> cv.fit.ri

Call:  cv.glmnet(x = X, y = y, nfolds = 10, alpha = c)

Measure: Mean-Squared Error

    Lambda  Measure         SE Nonzero
min  24653 1.018e+09 152895123       5
1se  27057 1.142e+09 193981597       4
> cv.fit.la          =     cv.glmnet(X, y, alpha = b, r
```

From the summaries on the left, Ridge Regression results in the lowest lambda for the usual rule and the lowest standard error.

Elastic-Net: compromise between Lasso and Ridge. It penalizes a mix of both absolute and squared size.

Lasso: penalizes the absolute size of coefficients. It offers automatic feature selection, because it can remove some features.

Ridge: Penalizes squared size of coefficients.
Ridge offers feature shrinkage.
Leads to smaller coefficients.



From the boxplots, we have random forest as the best model.

The other three tend to perform similarly to each other.

Time needed to train each model was the longest for the random forest.