Predicting Income and Unemployment Rate in the United States

Junchan Byeon, Alex Chen, David Chen, Suzanne Papik, Yinqi Zhang

Abstract

Among many factors, the two we deemed important to measuring overall health of individual counties were income and unemployment. By finding the significant predictors of those response variables, we can create a model to predict, with accuracy, how well off a certain county will be based on a given set of values. We will also be able to see how certain aspects of a certain county influence income and unemployment rate. To create our model, we used a 2015 American Community Survey from Kaggle. Because there are two different responses we are looking at, we will be doing a multiple linear regression on income and doing a logistic regression for unemployment rate. To calculate the best possible model for each variable, we used 10-fold cross validation for each alpha when looking at the LASSO, Ridge, and Elastic Net models. Based on a mean-squared error calculation for each model, we found that the elastic net model with alpha level 0.6 was the best fit for predicting income and the LASSO model was the best fit for predicting unemployment rate.

Keywords

Unemployment Rate
Income
General Linear Regression
Logistic Regression
American Census Survey

Introduction

Predicting the future of the economy has been an everlasting topic of study and research. Because determining the overall health of a county is complex and intertwined with other aspects of society as a whole, we decided to look at the strongest variables, income and unemployment, that are affected as a symptom of a growing or dying county. The US Census Bureau provides estimates every year which is denoted as the American Community Survey. Within Kaggle, we found a 2015 American Community Survey data set that we will be using to create separate models for both income and unemployment. In the data, there are 37 variables and 3221 observations. Each observation corresponds to a single county and the variables range from variables like race to mean commute time. By seeing which variables are large factors affecting the model, we would be able to provide a suggestion of which variable to target in order to help increase the response variable. For example, if something like public transportation had a large effect on the unemployment rate, we could recommend to a county that they should be devoting more funds towards public transportation. With the official US census coming up in 2020, we will be able to see how well our models are predicting each variable.

Methodology

Data Cleaning

In order to first understand what the data looks like and account for any potential blatant outliers, we plotted the income variable on a histogram and then on a scatter plot by state, demonstrated in Table 1 and Plot 1 respectively. While the histogram turned out to be skewed right, the scatterplot very clearly showed that there was a state that was a complete outlier compared to the rest - Puerto Rico. Thus, moving forward, we removed the observations since not only would it skew the entire results, but our focus is solely on the United States. We also discovered that there were two counties with unavailable data - Loving, Texas and Kalawao, Hawaii. Since both counties have extremely low populations, around 100 or less, they were removed from the data. Finally, the variable “women” was removed since it was entirely dependent on the population of men and total population, it is entirely unnecessary. Census ID, county, and state were also removed since we were just looking at the various predictors on a national level.

The data was split into a testing dataset and a training dataset, split 25-75 respectively. This was done so that we could create our model based on the training dataset and determine accuracy through the testing dataset.

General Linear Regression-Income

To start off, we first needed to check to make sure all the conditions were meant. Since the data is not linear, as shown by Plot 2, a box-cox transformation was conducted to determine the optimal lambda. A square root transformation of the income output was determined to be

ideal. Looking at the other conditions shown in Plot 4, there appears to be independence. While the data isn't necessarily normal according to the shapiro-wilks normality test, the dataset is sufficiently large that we can proceed forward with some caution. Data with extremely high influential points as determined by cook's distance were also looked at and removed. Most notably, there was only one point - Los Angeles, California (Plot 3) which needed to be removed.

Examining the VIF values in Table 2, it is clear that there are extreme collinearity issues. This was to be expected as most of the predictors are percent based, so as soon as one variable goes up the other has to go down. Thus, we first attempted to manually remove the predictors with large VIF values until they all are within a reasonable value. Predictors were then testing with ANOVA tables to determine if there were any significant variables that were removed.

The transformed and VIF-adjusted predictors were then put under a stepwise regression, using both AIC and BIC models. The best model was then picked from there and then predictions could be done based on the testing dataset to determine which was the ideal model in terms of predicting values.

LASSO/Ridge/Elastic Net Regression-Income

The methodology of tenfold cross-validation is employed to find the optimal alpha value. The training dataset is randomly split into 10 equal-sized subsets, and each of them will be used to fit a linear model with alpha values ranging from 0 to 1. In the analysis, we set the difference in between alpha values to be 0.1, so the alphas take value from 0, 0.1, 0.2, ..., to 1. For the training dataset, we removed the following predictors before fitting the model --- *CensusID, State, County, Women*, and *Income*. *CensusID* is for tracking use only, and factoring both *State* and *County* would introduce too many unwanted categorical predictors. The predictor *Women*, once included, would inflate vif and cause a potential collinearity problem.

If the alpha level is 0, the model fitted is Ridge regression; if the alpha level is 1, the model fitted is Lasso regression. Any decimal values in between represent a combination of Ridge and Lasso regressions. The underlying problem is that Ridge regression does not do variable selection while the Lasso regression does not deal with the collinearity issue. Another relatively innovative approach, namely the Elastic Net, is adopted to fit the model with alpha level 0.5. The Elastic Net approach minimizes both the sum of Beta j's and Beta j Squares.

$$(y - \mathbf{X}b)^\top (y - \mathbf{X}b) + \lambda\alpha \sum |b_j| + \lambda(1 - \alpha) \sum b_j^2.$$

Upon fitting ten different models, we first use them to predict the income using the testing dataset and then select the optimal alpha value by calculating the mean prediction errors. The mean prediction errors are either the means of squared difference between the actual income and the predicted ones or the means of the absolute value of the actual income and the predicted ones. The smaller the mean prediction error, the better the model is in terms of fitting the data points.

Logistic Regression-Unemployment

In order to predict the probability of being unemployed in a certain county, logistic regression was used. In order to be able to perform a logistic regression, first a binary variable had to be created from the dataset. Using data obtained from https://data.bls.gov/timeseries/LNS14000000, it was found that at the beginning of 2015, which was the year this American Community Survey was taken, the unemployment rate was 5.7%. Because of this, it was decided to create a variable that took on the value of 1 when the unemployment rate of a county was higher than the national unemployment rate, and 0 when then unemployment rate was lower than the national unemployment rate. From there, three models were fit on the dataset, which were LASSO, Ridge regression, and elastic net regression. This was done by using the training dataset, a logistic regression, and an alpha of 1, 0, and 0,5 respectively. After these models were fit, a cross validation was done for each of the three models. The solution paths for each of the variables and lambda were plotted following this to see the optimal lambda and which variables the models selected. In order to see how well the models fit the data, the next step was to predict new values using each model, and calculate the area under the ROC curve. This was done by using the predict function in R, using the cross validation fit, and the lambda that was selected from the cross validation. Once these predicted values from each of the tree models were generated, an ROC curve was plotted to see which model had the best predictive power. Since it was unable to be determined which one was best from the curves, the area under the curve had to be calculated. The curve with the highest area was determined to be the best model.

Discussion

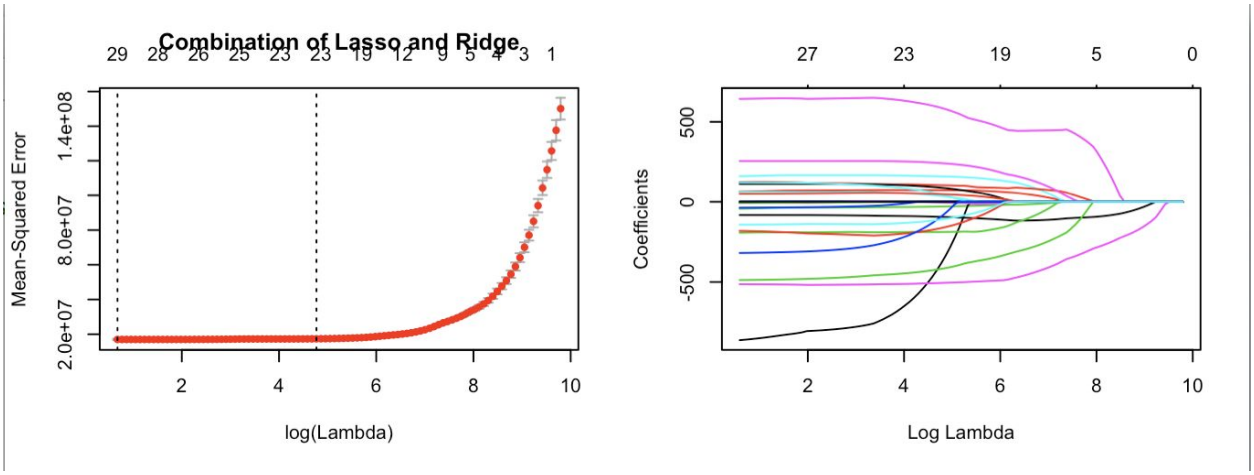
General Linear Regression-Income

Reducing the VIF values, there ended up being only 14 variables left out of the 31 original, as seen on Table 3. The AIC and BIC models both produced the model with 13 variables, with correlation coefficients shown on Table 4. Looking at their ability to predict, they had a mean error of $3171 and a mean squared error of 19660025. Considering the range of incomes and the overall size of the dataset, that mean error is relatively small and thus the model can be seen as relatively effective. Moving forward however, we needed to consider the different types of models that can be done.

LASSO/Ridge/Elastic Net Regression-Income

Among all the mean prediction errors, the smallest ones both correspond to alpha level 0.6. The mean prediction errors are 3,132.609 (absolute value) and 17,444,974 (difference squared). Therefore, the model with alpha level equal to 0.6 is the optimal model fitted. By calling the function $lambda.1se we acquire the lambda value, which is 117. 9132 in this case. The larger the shrinkage parameter lambda, the more strongly the coefficients are shrunk.

On the plot generated below, the left-hand-side is the cross-validation curve; the selected lambda levels (log scale) are marked by the two vertical lines. The right-hand-side portrays the variable fitting paths. As log lambda goes from 0 to 27, the number of predictors in the model increases.



Output 1 in appendix A is the variable selection process performed by the mix model. A dot implies that the predictor is not statistically significant and therefore should be removed. It is noticeable that the coefficients are largely shrunk to a 10^1 or 10^3 level. Due to the linear relation present, one example to interpret the coefficients is that with 1 percent increase in Hispanic population, there would be 71.303 dollars increase in local income associated. Another example would be with 1 percent increase in Carpool proportion, there would be 66.917 increase in local income associate. For future reference, the local government could encourage carpool or create more public-sector jobs if they want the local income to go up. It would also be of research interest to study the income distribution by different ethnicity groups and regions.

Logistic Regression-Unemployment

After the binary variable was created, the three models, LASSO, ridge regression, and elastic net regression, could be fit. The cross validation for each of the three models was also computed. From both of these, plots of each of the regression models and each of the cross-validation fits could be plotted in regards to the lambda. Plot 5 can be seen in appendix A showing the coefficient paths for each of the models, and the optimal lambdas for each model.

After these were plotted, the predicted yhat observation values were calculated. For each yhat, for the LASSO, ridge, and elastic net regression models, an ROC curve was constructed. A curve for each of the models can be seen in plots 6 through 8 in appendix A.

Each of the regression models seem to have very good predictive power by looking at the curves and how quickly they rise. In order to determine which model actually was best, the area under the three curves was calculated. The AUC (area under the curve) for the LASSO model was **0.9070913**, the AUC for the ridge model was **0.903377**, and the AUC for the elastic net model was **0.9048422**. The goal is for the AUC to be as close to 1 as possible, so it can be seen that each of the models do a phenomenal job at predicting. LASSO is the largest value by several thousandths. Because of this, LASSO was determined to be the best model for predicting if a county will be above or below the national unemployment rate.

When we look at the LASSO model, it can be seen that several different predictors were taken out of the model. Eight predictors were taken out, and twenty-four variables, plus the intercept, remained. To see all of the predictors included in the model, see output 2 in appendix A.

Conclusion

Looking at the income predictions, the elastic net proved to be the most accurate out of any of the other models, selecting 21 variables. Looking at the correlation coefficients, the most significant positive predictors were the percent of asian populations in a county as well as the mean commute time to work while the most negative predictors were poverty (including child poverty), the percent of workforce that is self employed, and the total number of citizens.. While we can't draw definitive conclusions about causation, a possible reasoning that will need to be tested is that areas with higher mean commute time have greater potential to work in multiple areas or that they'll be compensated a higher income and thus the income goes up. Also, that potentially immigrants tend to target areas with higher incomes and live there. For the negative predictors, it is reasonable to assume that higher levels of poverty indicate lower incomes as the definition of poverty comes from income. The percent of self employed also tells a similar story as it is possible that rather than being an indicator of innovation and entrepreneurship, it could be an indication of a lack of jobs and forcing citizens into other areas in the workforce. The number of citizens could just be an indicator that larger areas such as cities have a greater presence of the poor.

For the unemployment predictors, there were a total of 24 variables after conducting a lasso regression -- the best possible model in terms of predictions. In terms of predicting whether or not a county would have an unemployment level higher than the national average of 5.7%, citizen population and income per capita were the biggest positive predictors while Income was the largest negative predictor. This suggests that larger populations have a higher probability of unemployment as well as the income per capita. The income per capita being a strong positive predictor is something that will need to be investigated as it seems to be counterintuitive. The Income being the largest negative predictor makes sense as counties with higher median incomes will likely be in areas where unemployment is lower.

Looking at the differences between the two models in terms of which predictors affect the response variable, there is a lot of overlap. This is a positive aspect of what we are trying to accomplish with this group project. If we are looking at how to evaluate overall health of individual counties and trying to look which factors have an effect on the health, if our models overlap, it is likely that both of those response variables are good indicators of health of a county. The two noticeable differences between the models is the predictors: Walk and Other Transportation which correspond to the percentage that walk to work and the percentage that use other transportation to get to work. These predictors are significant in the unemployment model while not significant in the income model. This makes sense because the unemployment rate most likely will be affected by the percentage of people who are walking or taking another transportation to work. If there is a large percentage walking to work, there will probably be fewer people unemployed.

As always, there is a lot of work that could be done in order to improve this model or go beyond what this model tells us. One example to improve this model is to look at changing the percentage for which we are splitting the data. Usually, training on 75%-25% model is inferior to training on a 90%-10% model. By training on more data, the model should be more reliable and have a better accuracy. Another thing we could do to improve accuracy of choosing which alpha to select is using bootstrapping. By taking a different training set multiple times, we could use this to take an average of the mean squared error. This is helpful in case we are accidentally doing a poor split of the data. This would increase the likelihood of using the correct alpha. Future work that goes beyond what we have done could include using some separate data set that would rank the counties from 1 to 3221 based on how good it is to live in that county. Instead of using unsure indicators of health like income and unemployment rate, this would allow us to use that variable as our response and help us make a really good model to predict the health of a county.

References

MuonNeutrino. “US Census Demographic Data | Kaggle.” Countries of the World | Kaggle, Kaggle, 14 Dec. 2017, www.kaggle.com/muonneutrino/us-census-demographic-data/data.

Appendix A

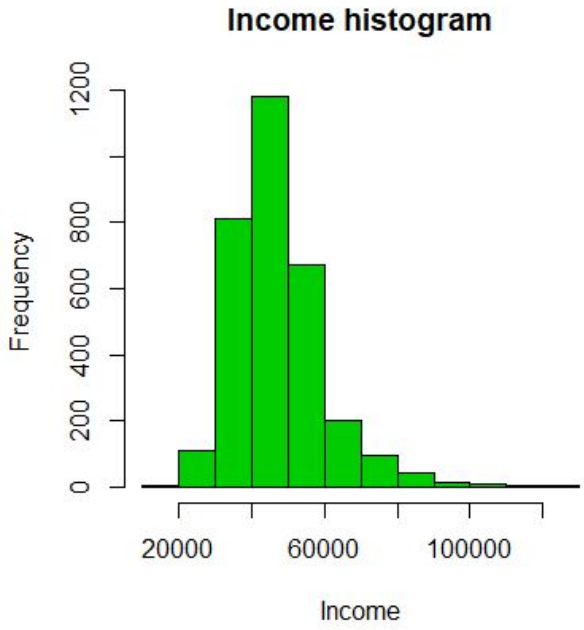
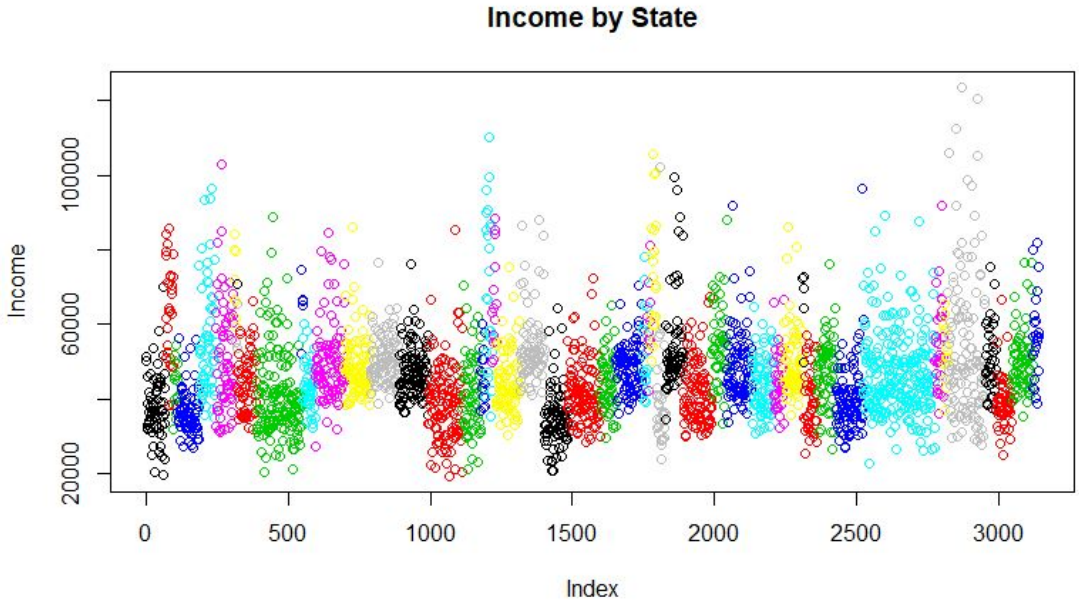


Table 1



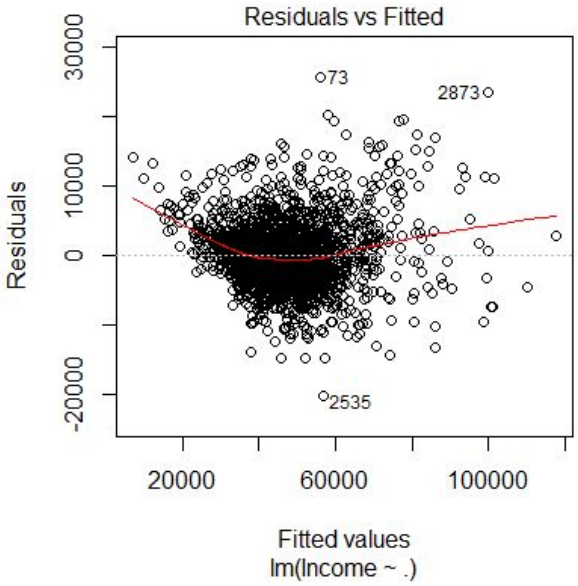
Plot 1

```
> vif(lm_allt)
      TotalPop            Men       Hispanic          White
  7451.110911    6464.941914     183.156687     266.644104
          Black         Native          Asian        Pacific
   102.734727      31.231106       6.858992       1.858937
        Citizen      IncomeErr  IncomePerCap IncomePerCapErr
   205.375209       2.332470       5.922559       2.403603
        Poverty    ChildPoverty Professional        Service
    13.696823       9.911041   10078.101624    3283.472130
         Office   Construction   Production          Drive
  2539.436610    4408.171850    8183.775501   12253.759251
        Carpool        Transit           Walk     OtherTransp
  1788.326863    1976.091792    2892.623242     578.205941
     WorkAtHome  MeanCommute       Employed    PrivateWork
  2129.373124       1.572328     278.635023   19203.702994
      PublicWork  SelfEmployed     FamilyWork   Unemployment
 13105.435953    4808.750795      65.538771       2.746457
```

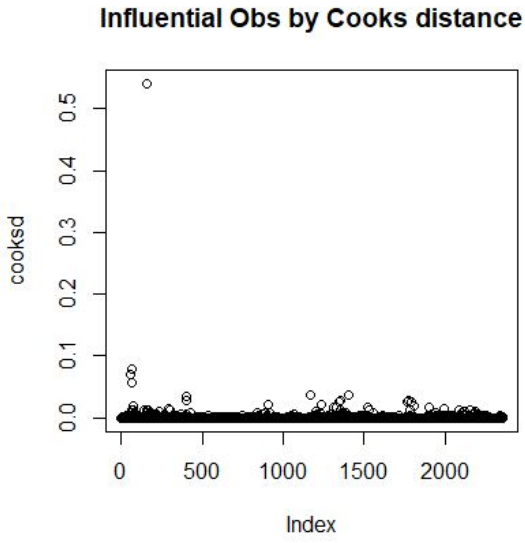
Table 2

```
vif(lm_manual)
  TotalPop        White        Black       Native        Asian
  1.431200     2.720177     2.576476     1.653097     2.123701
   Pacific IncomePerCap      Poverty Professional        Drive
  1.283367     4.443992     3.580224     2.820479     2.043680
WorkAtHome  MeanCommute  PrivateWork Unemployment
  2.091952     1.290788     1.975467     2.208786
```

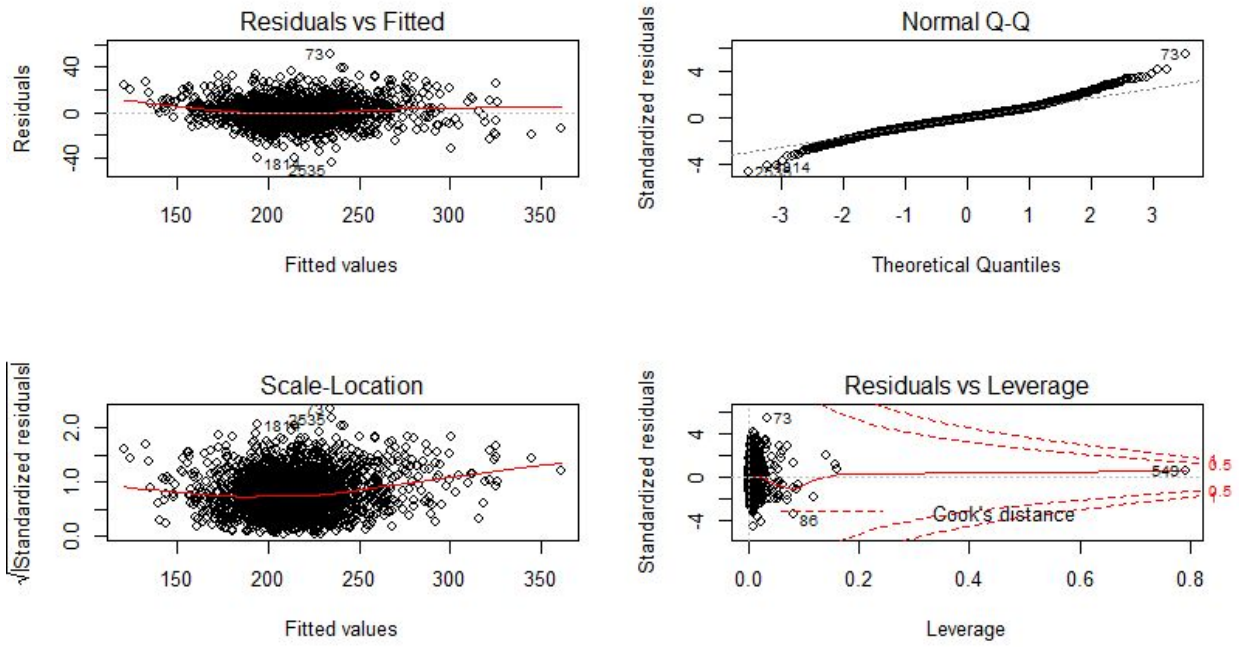
Table 3


Residuals vs Fitted
73
2873
2535
Residuals
30000
10000
0
-20000
20000
60000
100000
Fitted values
lm(Income ~ .)

Plot 2


Influential Obs by Cooks distance
cooksd
0.0
0.1
0.2
0.3
0.4
0.5
0
500
1000
1500
2000
Index

Plot 3



Plot 4

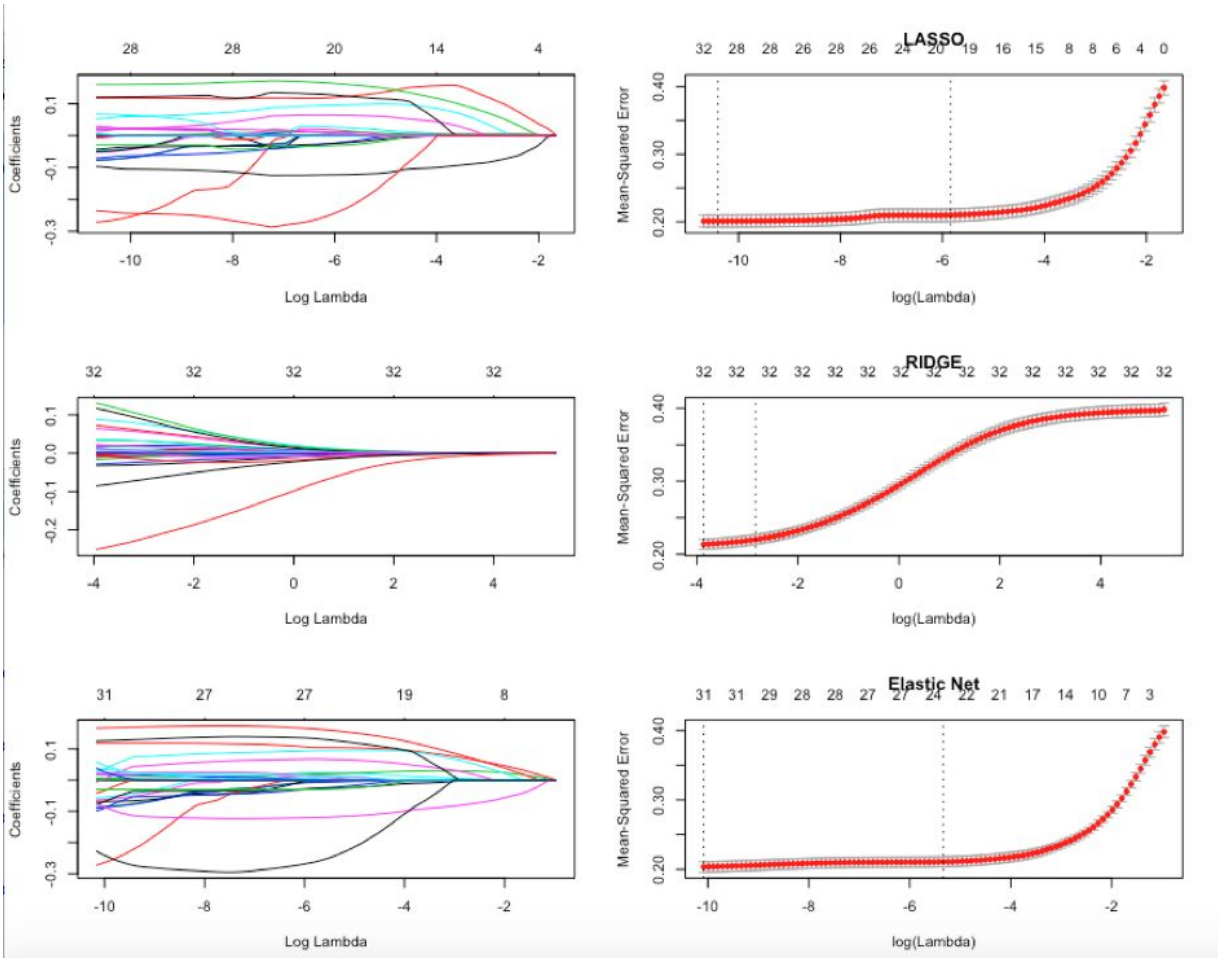
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.641e+02  3.763e+00  43.618  < 2e-16 ***
TotalPop     -5.636e-06  9.780e-07  -5.763 9.36e-09 ***
White        -2.861e-01  1.676e-02 -17.070  < 2e-16 ***
Black        -1.948e-01  2.196e-02  -8.868  < 2e-16 ***
Native        3.678e-01  3.709e-02   9.917  < 2e-16 ***
Asian         9.716e-01  1.085e-01   8.953  < 2e-16 ***
Pacific      -5.640e-01  2.661e-01  -2.120  0.03413 *
IncomePerCap  2.160e-03  7.023e-05  30.751  < 2e-16 ***
Poverty      -1.843e+00  5.677e-02 -32.463  < 2e-16 ***
Professional  4.734e-01  5.132e-02   9.224  < 2e-16 ***
WorkAtHome   -4.268e-01  7.704e-02  -5.540 3.36e-08 ***
MeanCommute   5.848e-01  4.037e-02  14.484  < 2e-16 ***
PrivateWork   3.642e-01  3.598e-02  10.123  < 2e-16 ***
Unemployment -2.320e-01  8.378e-02  -2.769  0.00567 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.531 on 2342 degrees of freedom
Multiple R-squared:  0.8802,    Adjusted R-squared:  0.8795
F-statistic:  1324 on 13 and 2342 DF,  p-value: < 2.2e-16
```

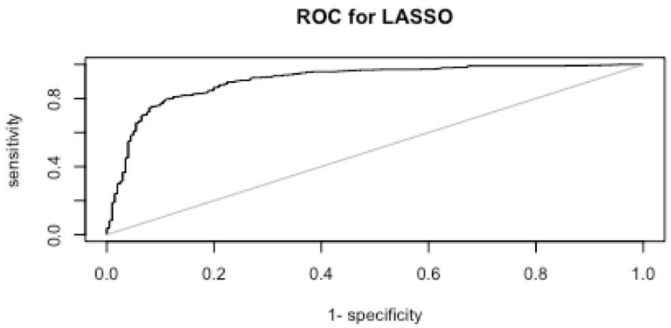
Table 4

```
(Intercept)      2.767746e+04
TotalPop          .
Men               .
Hispanic          7.130313e+01
White            -2.642650e+01
Black             .
Native            1.528397e+02
Asian             5.238716e+02
Pacific          -9.184375e+00
Citizen          -1.316442e-03
IncomeErr         2.811070e-01
IncomePerCap      1.181609e+00
IncomePerCapErr  -1.088883e+00
Poverty          -4.964837e+02
ChildPoverty     -1.041434e+02
Professional      1.002065e+02
Service          -1.888259e+02
Office            .
Construction      1.054030e+01
Production       -1.322014e+01
Drive             .
Carpool           6.691704e+01
Transit          -1.106874e+02
Walk              .
OtherTransp       .
WorkAtHome        .
MeanCommute       2.258283e+02
Employed          .
PrivateWork       .
PublicWork        3.672546e+01
SelfEmployed     -3.794166e+02
FamilyWork        .
Unemployment     -8.190530e+01
```

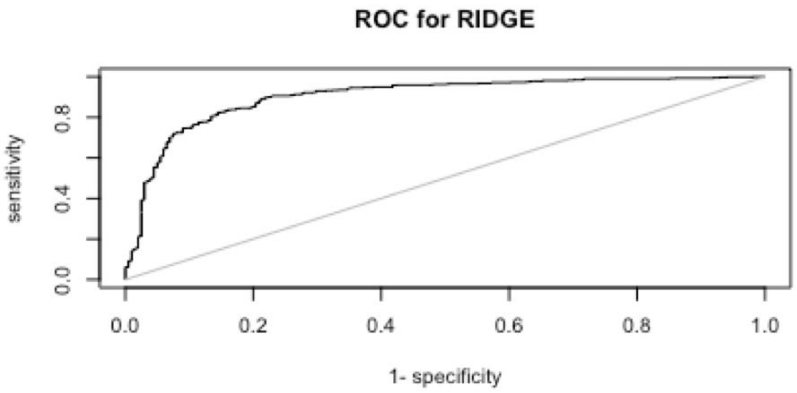
Output 1


LASSO
RIDGE
Elastic Net
Coefficients
Log Lambda
Mean-Squared Error
log(Lambda)

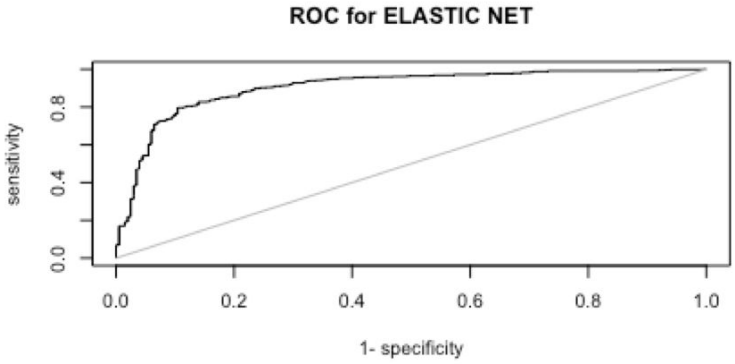
Plot 5


ROC for LASSO
sensitivity
1- specificity

Plot 6


ROC for RIDGE
sensitivity
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
1- specificity

Plot 7


ROC for ELASTIC NET
sensitivity
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
1- specificity

Plot 8

```
(Intercept)      9.399631e-01
TotalPop          .
Men               .
Hispanic        -1.470381e-02
White           -1.642101e-02
Black            1.727910e-02
Native           7.595874e-03
Asian           -2.301892e-02
Pacific           .
Citizen          5.820210e-06
Income          -6.375330e-05
IncomeErr       -8.213828e-05
IncomePerCap     2.067713e-05
IncomePerCapErr -1.418053e-04
Poverty          1.184282e-01
ChildPoverty     1.116098e-02
Professional    -3.714666e-02
Service          9.031317e-02
Office           6.369201e-02
Construction    -3.027857e-02
Production        .
Drive           -3.209371e-02
Carpool          3.335307e-03
Transit           .
Walk             1.633950e-02
OtherTransp      1.345186e-01
WorkAtHome        .
MeanCommute      1.710650e-01
Employed          .
PrivateWork       .
PublicWork       1.615714e-02
SelfEmployed    -1.240718e-01
FamilyWork      -2.772516e-01
```

Output 2

Appendix B

Group Meetings

4/9-Meeting with the professor to determine what to do with the dataset

4/16-Meeting with the professor to determine what to do with high collinearity issues and to make sure we are on track with the project

4/19 - Exploratory meeting, all met in the library to determine how to split up the project, figure out the goals of the project, and start working on it

4/22-Met in the statistics department to continue working on the presentation and report. Finished working on the coding aspect of the project and made final determinations on models. Continued finishing up the presentation

4/23-Meeting with the professor to discuss how to interpret solution paths and regression models for median income and unemployment rate

4/24- Group video chat in order to practice the presentation, determine who was presenting what, and all be on the same page

Individual Contributions

Alex Chen - Tableau (visualizations), Abstract/Introduction/Conclusion, Introduction and Conclusion of the presentation
22.5%

David Chen - Exploratory Analysis, Conditions, General Linear Regression, Conclusion, Income GLR and exploratory analysis for the presentation
22.5%

Junchan Byeon - Abstract, Introduction, Conclusion
10%

Suzanne Papik - Unemployment Logistic Regression methodology and discussion/General Report Organization/Keywords, Unemployment Logistic Regression for the presentation
22.5%

Yinqi Zhang - Lasso/Ridge/Elastic Net Regression (Methodology and Analysis), Income LASSO/Ridge/Elastic Net for the presentation
22.5%