Predicting Income and Employment in the US

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

AMERICAN COMMUNITY SURVEY

U.S. CENSUS BUREAU

Motivation

- Determine factors that predict Income/Unemployment

Procedure

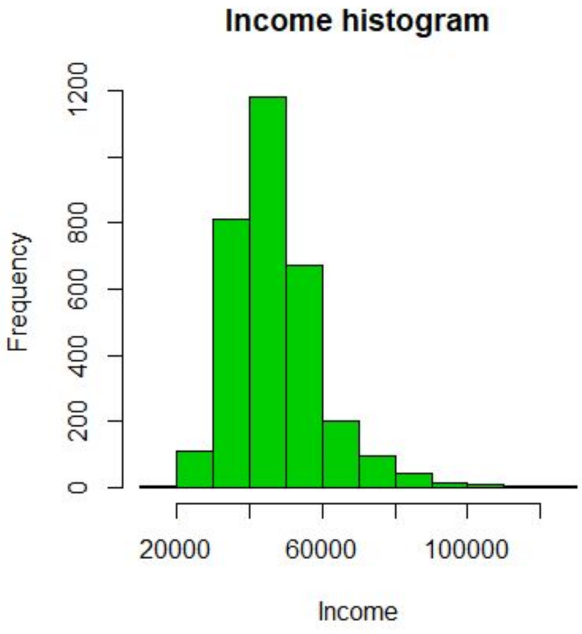
- Data Cleaning
- Exploratory Analysis
- Variable Selection
- Data Modeling
- Visualization
- Analysis
- Conclusion

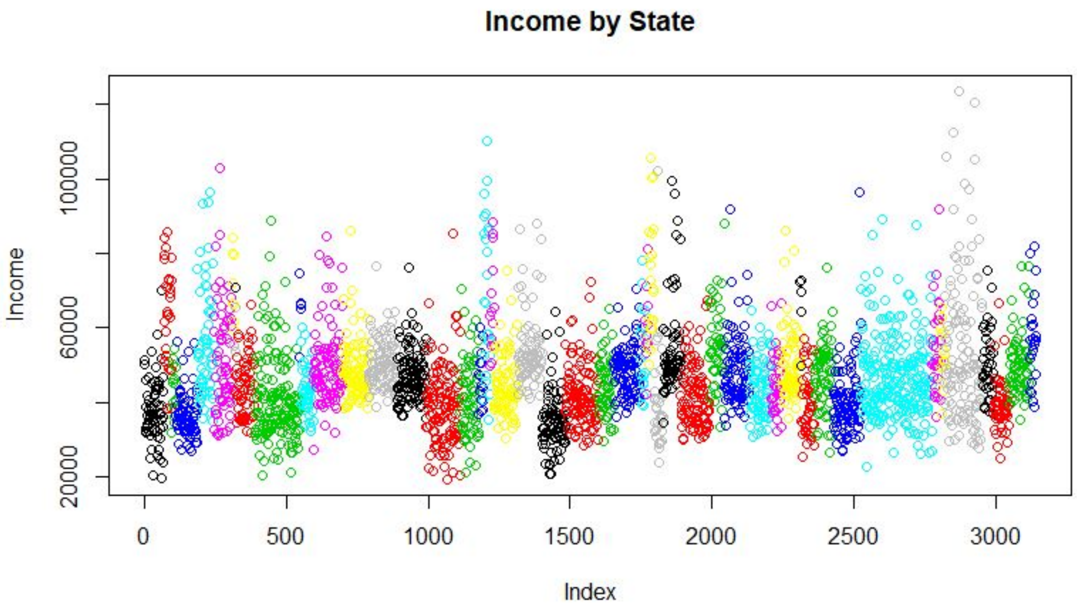
Data Cleaning

1. Remove NAs (***)
2. Training and Testing datasets (75-25)
3. Removed the Census Id, State, County
4. Remove Puerto Rico
5. Dependent variables

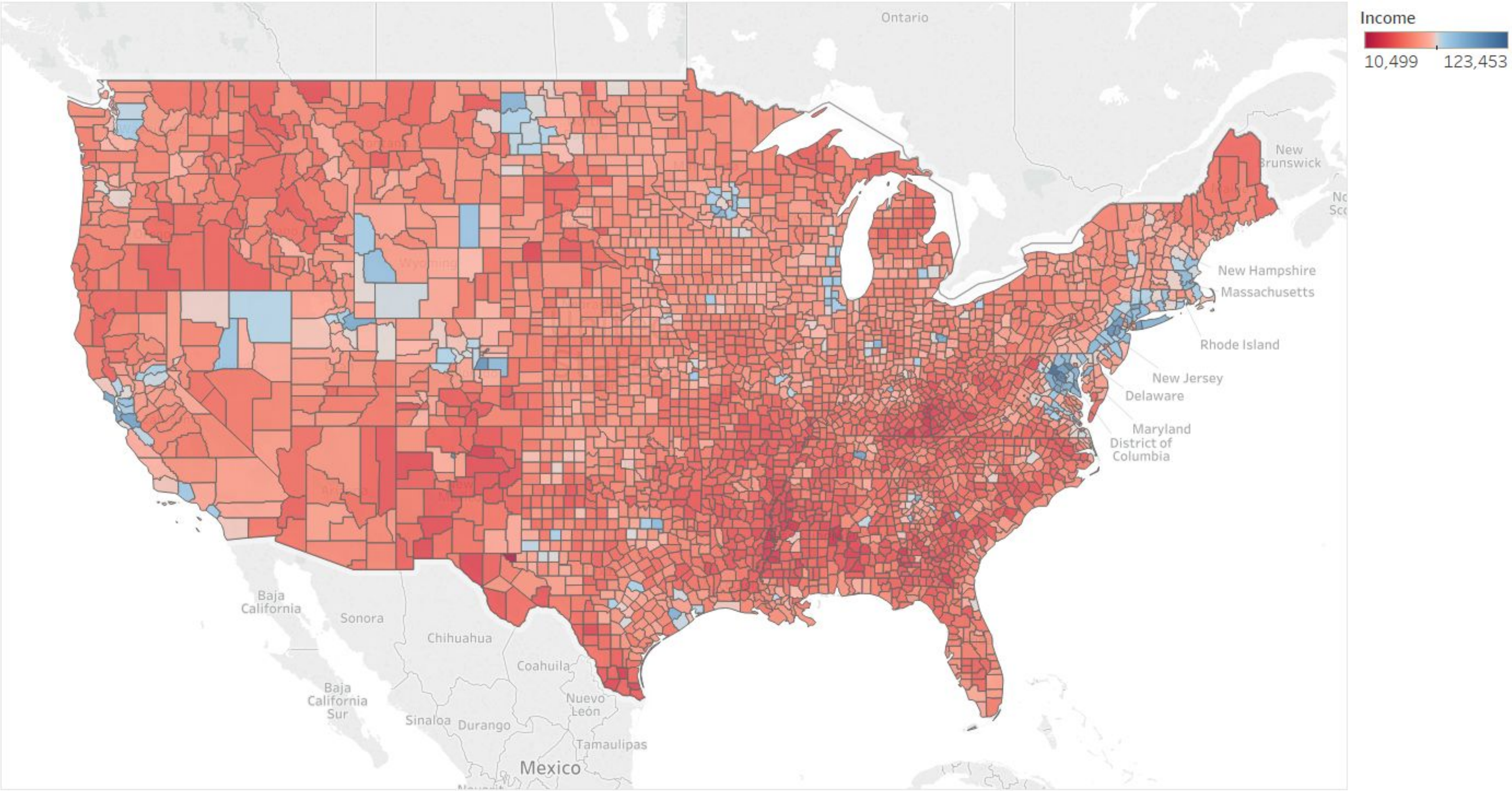
CensusId	State	County	TotalPop	Men	Women	Hispanic	White	Black	Native
1001	Alabama	Autauga	55221	26745	28476	2.6	75.8	18.5	0.4
1003	Alabama	Baldwin	195121	95314	99807	4.5	83.1	9.5	0.6
1005	Alabama	Barbour	26932	14497	12435	4.6	46.2	46.7	0.2
1007	Alabama	Bibb	22604	12073	10531	2.2	74.5	21.4	0.4

Exploratory Data Analysis

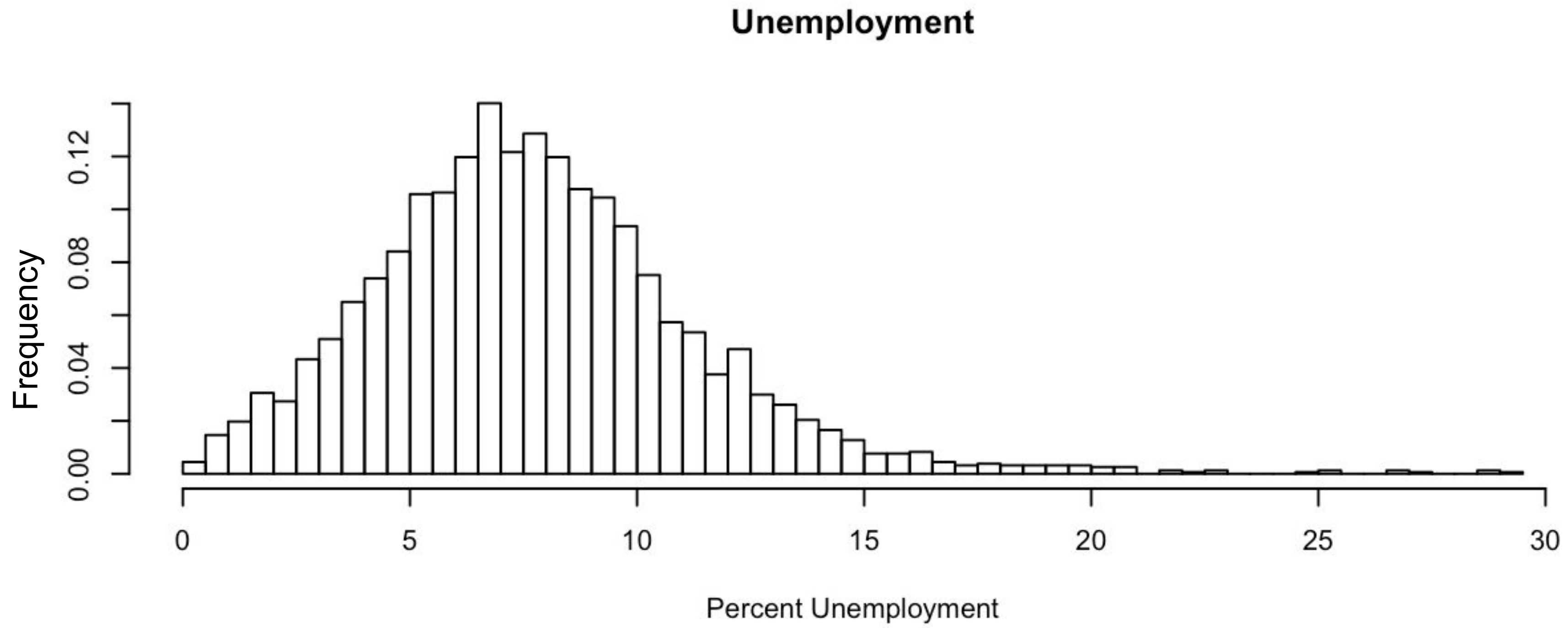


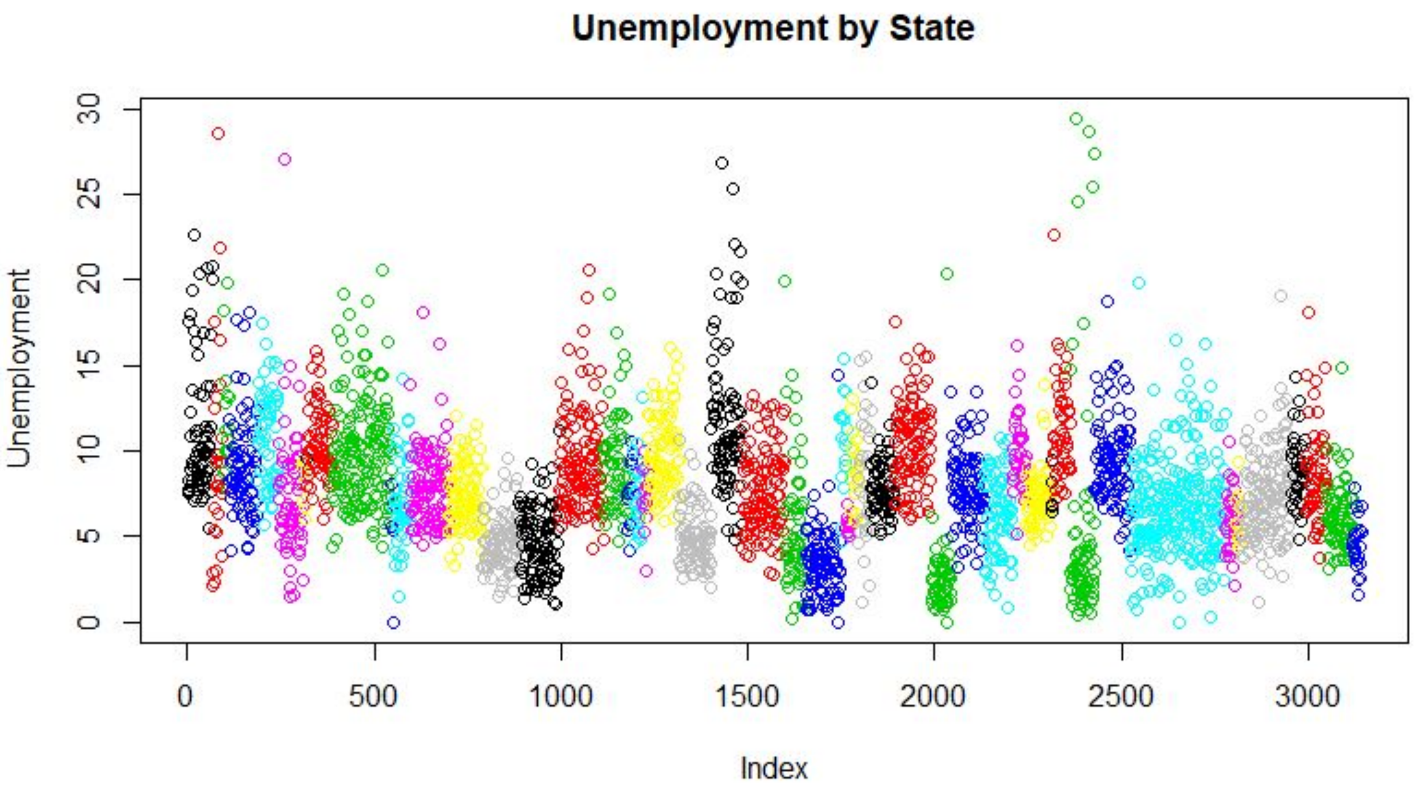


Income Map of United States

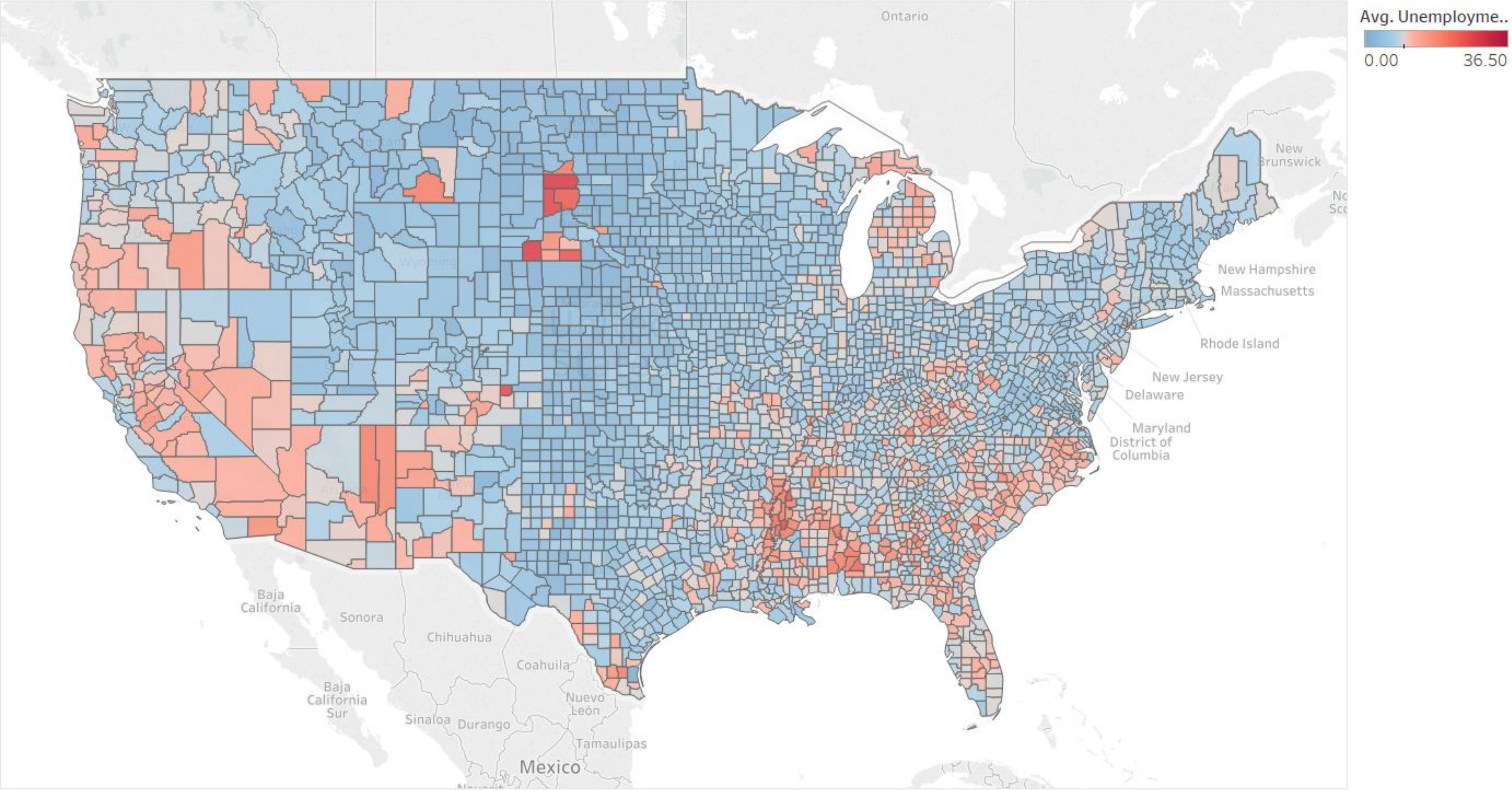


Map based on Longitude (generated) and Latitude (generated). Color shows sum of Income. Details are shown for State and County.


Unemployment
Frequency
0.00
0.04
0.08
0.12
0
5
10
15
20
25
30
Percent Unemployment


Unemployment by State
Unemployment
0
5
10
15
20
25
30
Index
0
500
1000
1500
2000
2500
3000

Unemployment by County



Map based on Longitude (generated) and Latitude (generated). Color shows average of Unemployment. Details are shown for State and County.

Variable Selection

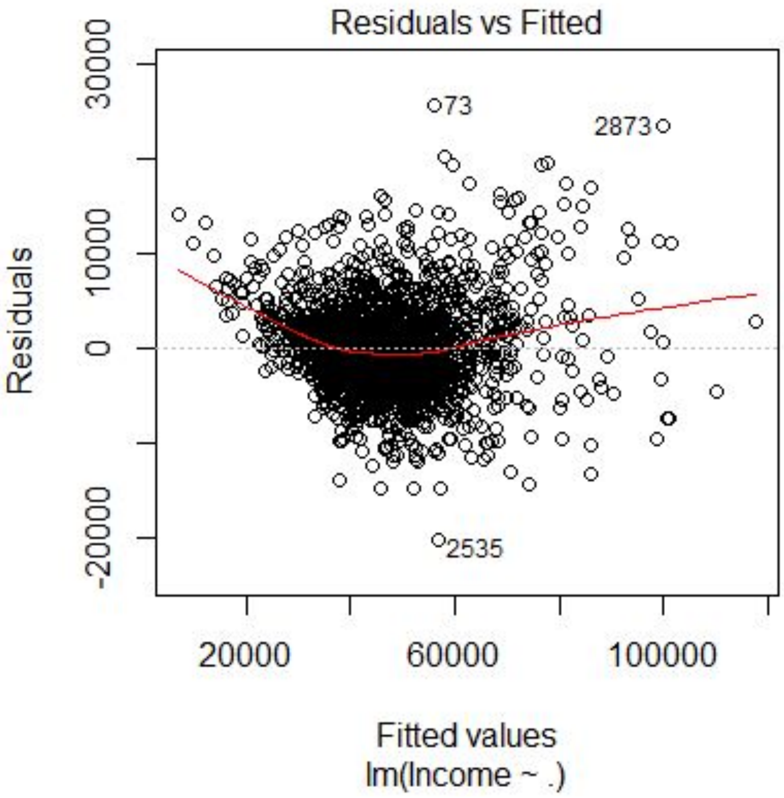
```
> vif(lm_allt)
     TotalPop             Men        Hispanic           White
  7451.110911     6464.941914      183.156687      266.644104
        Black          Native           Asian         Pacific
   102.734727       31.231106        6.858992        1.858937
      Citizen       IncomeErr    IncomePerCap 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
```

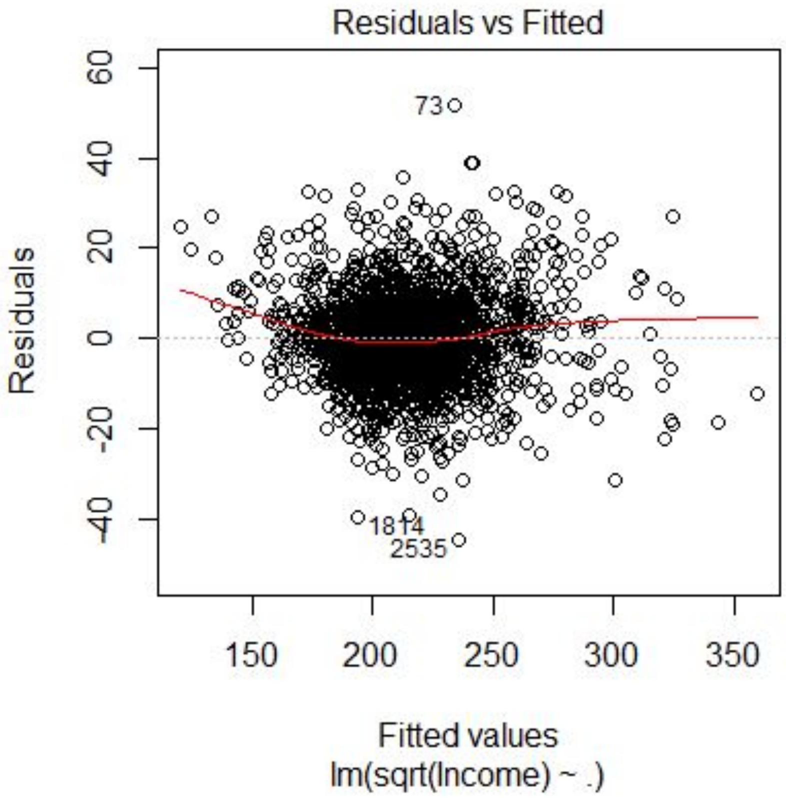
Multiple Linear Regression - Income

Data Modeling-Income

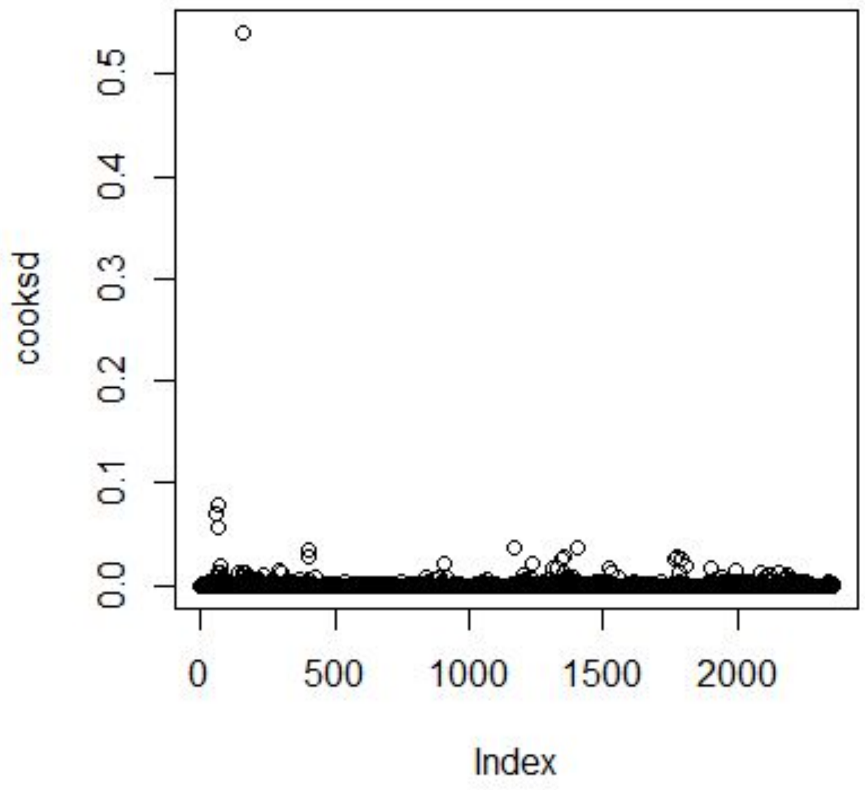
```
vif(lm_manual)
   TotalPop        White        Black       Native        Asian
   1.431200     2.720177     2.576476     1.653097     2.123701
    Pacific IncomePerCap      Poverty Professional        Drive
   1.283367     4.443992     3.580224     2.820479     2.043680
 WorkAtHome  MeanCommute 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

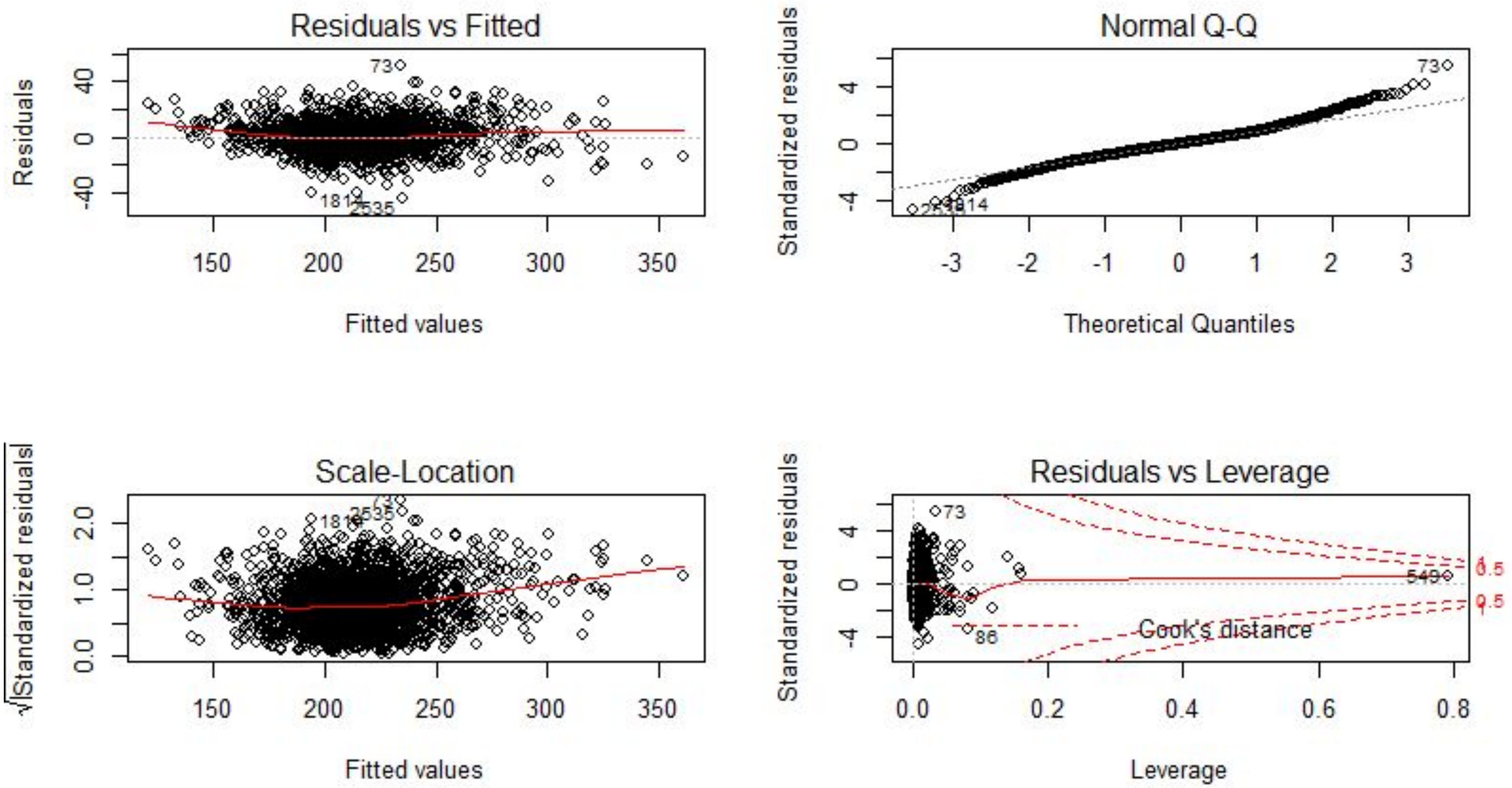
Transforming the Data





Influential Obs by Cooks distance




Residuals vs Fitted
Residuals
Fitted values
73
Normal Q-Q
Standardized residuals
Theoretical Quantiles
73
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
73
86
549
Cook's distance
0.5
1

AIC/BIC Model

```
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
```

Data Modeling-Income

Fit the LASSO, Ridge, and Elastic Net models:

```
fit.lasso<-glmnet(x.train,y.train,family='gaussian',alpha=1)
fit.ridge<-glmnet(x.train,y.train,family='gaussian',alpha=0)
fit.elnet<-glmnet(x.train,y.train,family='gaussian',alpha=0.5)
```

Creates 10-fold Cross Validation for each alpha:

```
for (i in 0:10) {
  assign(paste('fit',i,sep=''),
         cv.glmnet(x.train,y.train,type.measure='mse',alpha=i/10,family='gaussian'))
}
```

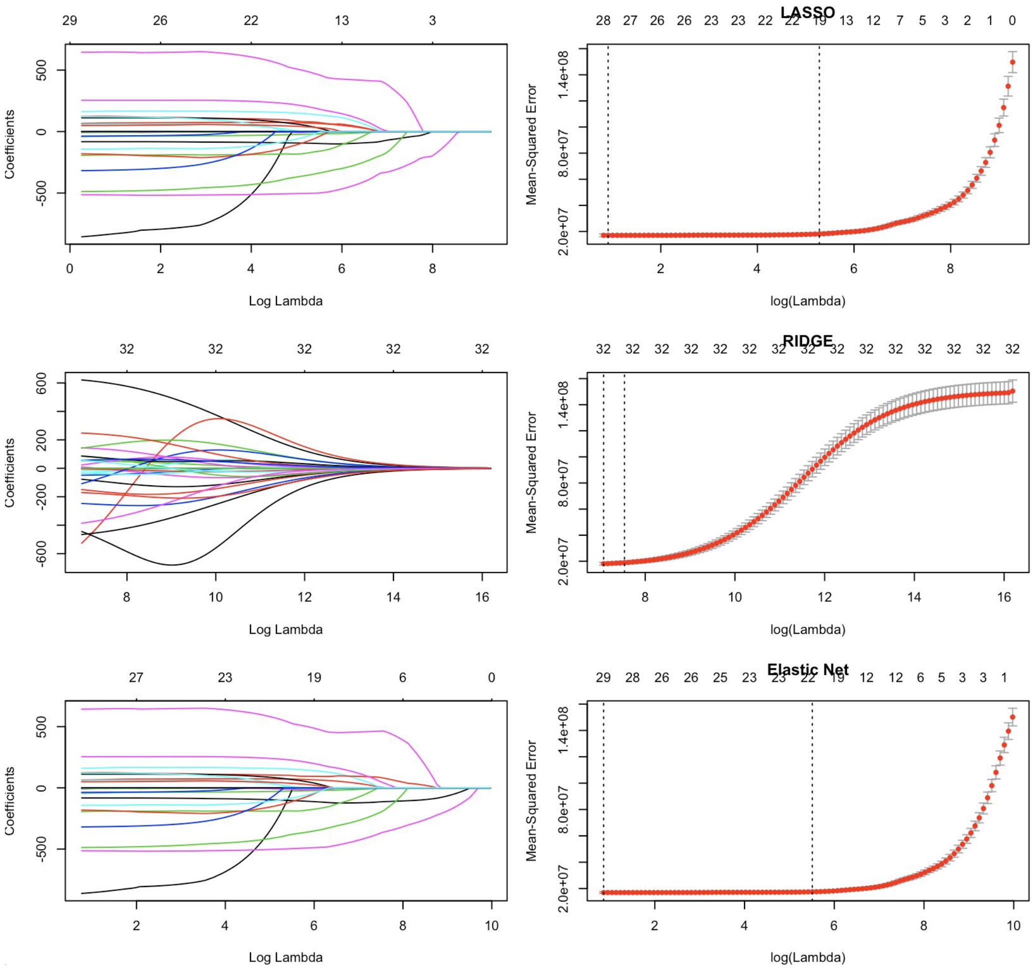
Plot the solution path and cross-validated MSE as function of λ

```
plot(fit.lasso,xvar='lambda')
plot(fit10,main='LASSO')

plot(fit.ridge,xvar='lambda')
plot(fit0,main='RIDGE')

plot(fit.elnet,xvar='lambda')
plot(fit5,main='Elastic Net')
```

LASSO

RIDGE

ELASTIC NET



Prediction-Income

Predict yhat0 to yhat10 using the fit for each alpha

```
yhat0<-predict(fit0,s=fit0$lambda.1se,newx=x.test)
```

Compute the Mean Absolute Error and Mean Square Error for each yhat

```
(mean(abs(y.test-yhat0)))
```

```
(mse0<-mean((y.test-yhat0)^2))
```

Fitting the Income Model

```
fit.AIC.BIC <- step(lm_manual2, direction = "both", k = 1, trace = 0)
```

MAE= 3171 **MSE=19660025**

```
fit.lasso<-glmnet(x.train,y.train,family='gaussian',alpha=1)
```

MAE= 3187.912 **MSE=18104664**

```
fit.ridge<-glmnet(x.train,y.train,family='gaussian',alpha=0)
```

MAE= 43654.52 **MSE=19009876**

```
fit.elnet<-glmnet(x.train,y.train,family='gaussian',alpha=0.5)
```

MAE= 3152.585 **MSE=17675860**

Elastic Net

- Reduce VIF
- Remove Insignificant Predictors

(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

Logistic Regression - Unemployment

Data Manipulation

- Had to create new binary variable in the dataset
- National unemployment rate in January of 2015 was 5.7%
- Created a binary variable that took the value 1 when the unemployment rate was greater than or equal to 5.7, and 0 when the unemployment rate was less than 5.7

```
cendata$unemploy<-ifelse(cendata$Unemployment>=5.7,cendata$unemploy<-1,cendata$unemploy<-0)
```

Employed	PrivateWork	PublicWork	SelfEmployed	FamilyWork	Unemployment	unemploy
2838	68.9	26.0	5.1	0.0	20.8	1
8894	74.3	16.0	9.6	0.1	9.6	1
2519	78.6	15.4	5.9	0.2	2.9	0
3787	77.9	18.9	3.2	0.0	2.1	0
152355	73.3	20.9	5.7	0.1	6.7	1

Data Modeling-Unemployment

Fit the LASSO, Ridge, and Elastic Net models:

```
fit.lasso2<-glmnet(x.train2,y.train2,family='binomial',alpha=1)
fit.ridge2<-glmnet(x.train2,y.train2,family="binomial",alpha=0)
fit.elnet2<-glmnet(x.train2,y.train2,family='binomial',alpha=0.5)
```

Creates 10-fold Cross Validation for each alpha:

```
for (i in 0:10) {
  assign(paste('fit',i,sep=''),cv.glmnet(x.train2,y.train2,type.measure='mse',alpha=i/10,family='binomial'))
}
```

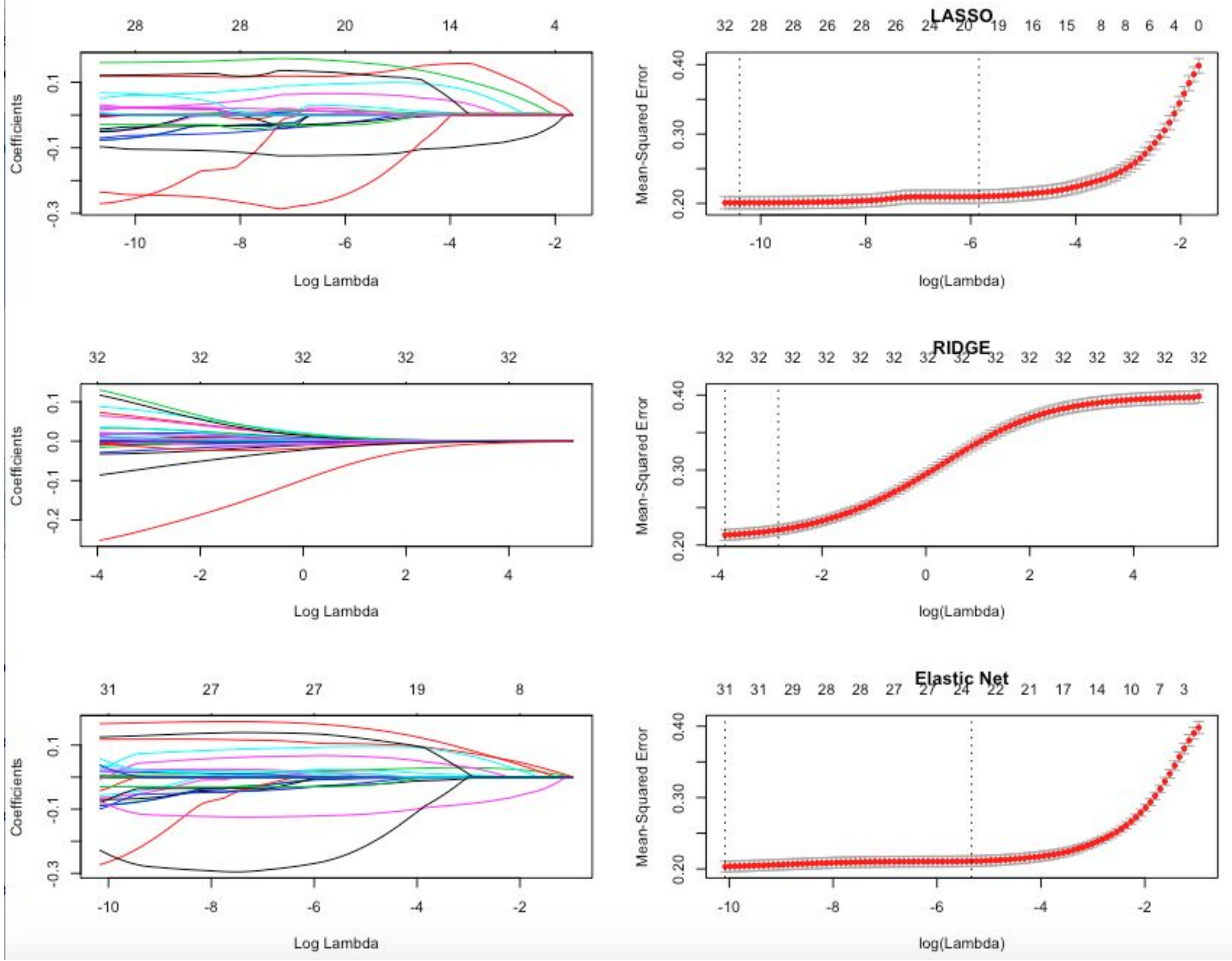
Plot the solution path and cross-validated MSE as function of λ

```
plot(fit.lasso2,xvar='lambda')
plot(fit10,main='LASSO')

plot(fit.ridge2,xvar='lambda')
plot(fit0,main='RIDGE')

plot(fit.elnet2,xvar='lambda')
plot(fit5,main='Elastic Net')
```

LASSO

RIDGE

ELASTIC NET



Prediction-Unemployment

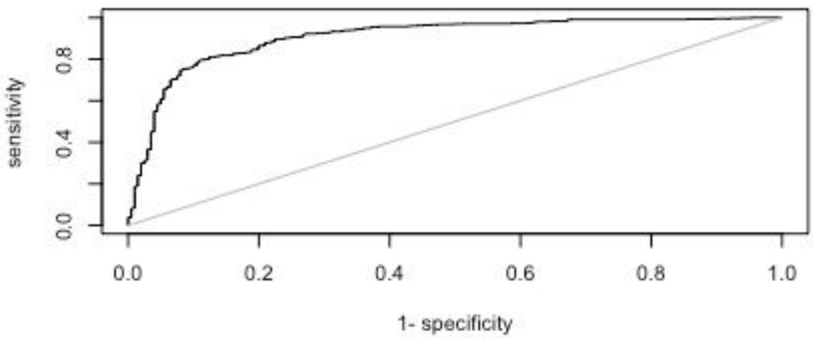
Predict yhat0 to yhat10 using the fit for each alpha

```
yhat0.2<-predict(fit0,s=fit0$lambda.1se,newx=x.test2)
```

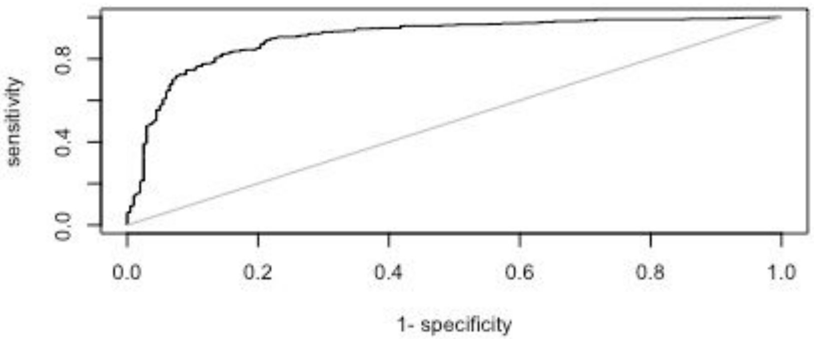
Compute The ROC curve and AUC for each model

```
roc.res0=roc(yhat0.2,factor(y.test2))   #Ridge
auc(roc.res0)
roc.res5=roc(yhat5.2,factor(y.test2))   #Elnet
auc(roc.res5)
roc.res10=roc(yhat10.2,factor(y.test2)) #LASSO
auc(roc.res10)
```

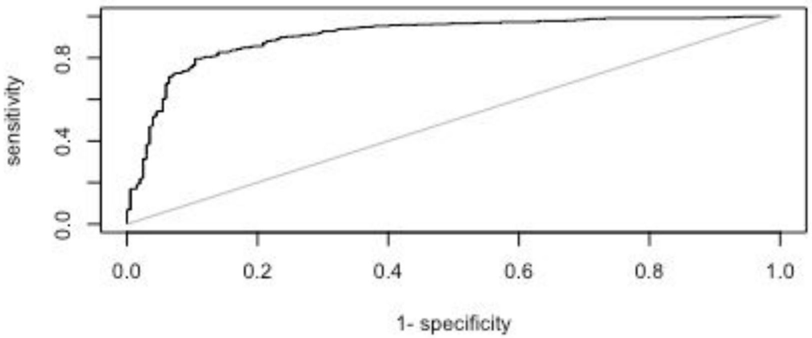
ROC for LASSO



ROC for RIDGE



ROC for ELASTIC NET



Fitting the Unemployment Model

fit.lasso2<-glmnet(x.train2,y.train2,family='binomial',alpha=1)

AUC=0.9070913

fit.ridge2<-glmnet(x.train2,y.train2,family="binomial",alpha=0)

AUC=0.903377

fit.elnet2<-glmnet(x.train2,y.train2,family='binomial',alpha=0.5)

AUC=0.9048422

LASSO Regression

- Removes insignificant predictors
- Shrinks insignificant predictors to 0

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

Final Models

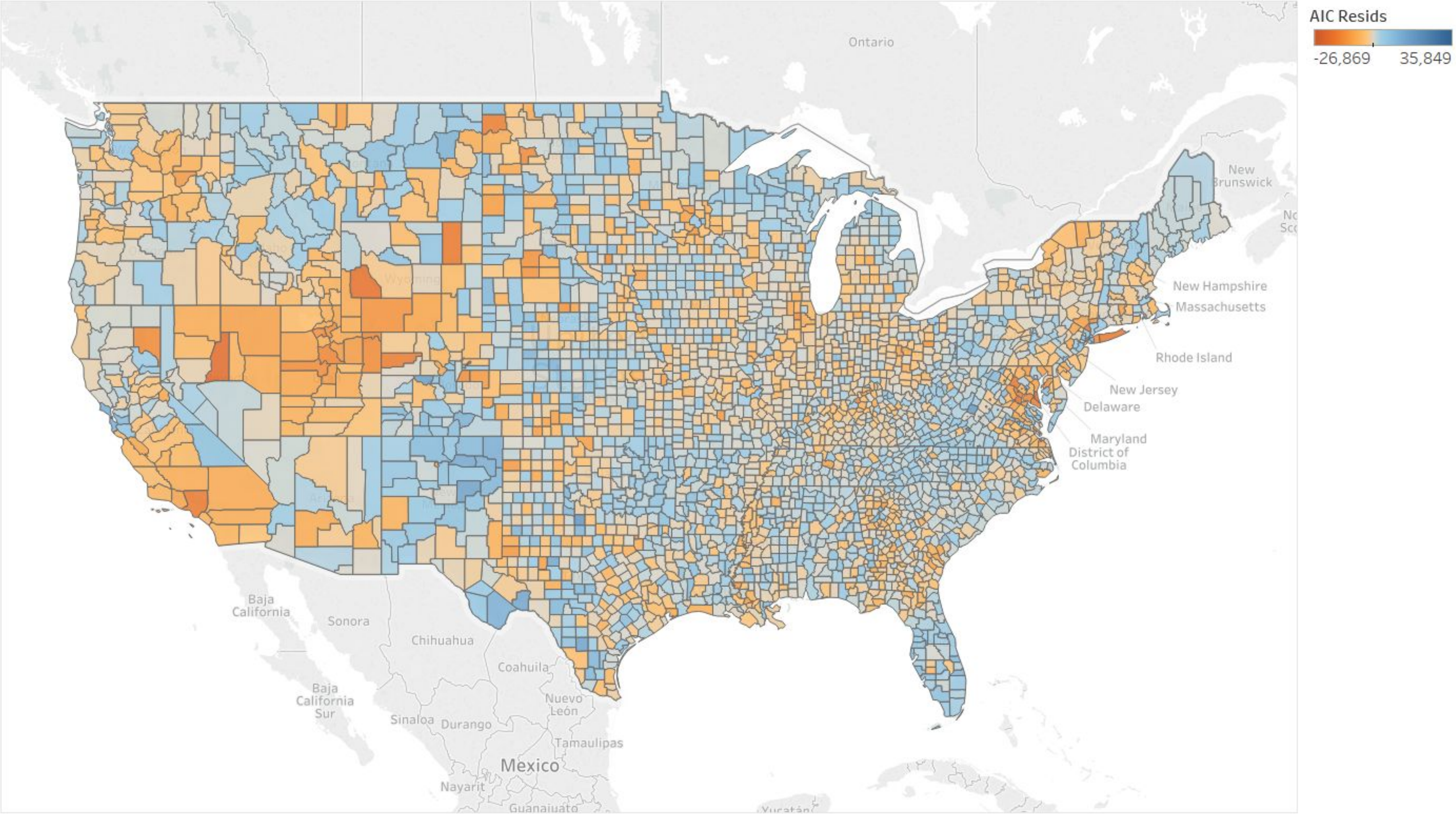
Model for predicting income(Elastic net)

```
fit.elnet<-glmnet(x.train,y.train,family='gaussian',alpha=0.5)
```

Model for predicting unemployment rate(Lasso)

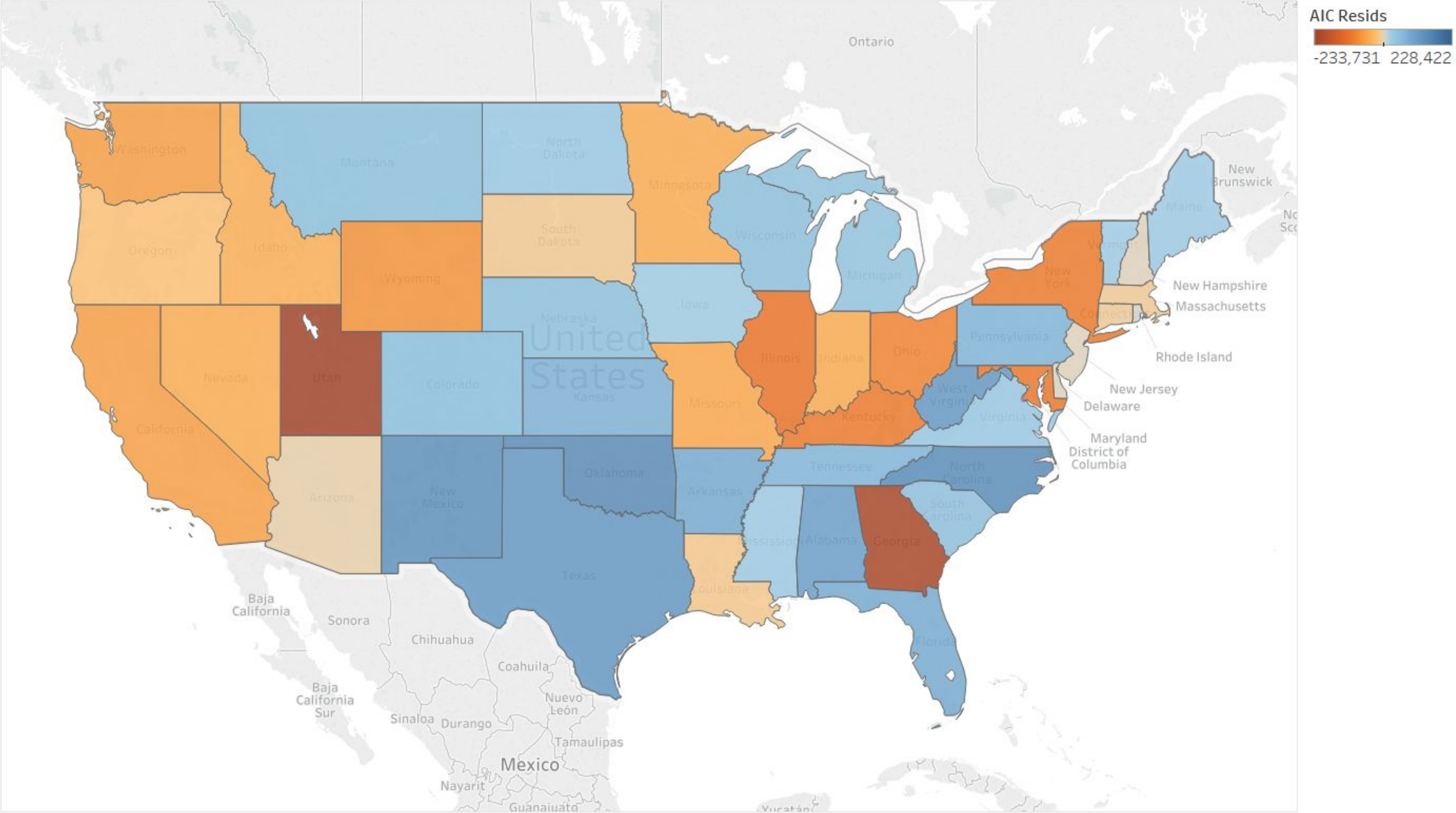
```
fit.lasso2<-glmnet(x.train2,y.train2,family='binomial',alpha=1)
```

Residual Heat Map through Forward/Backward Selection



Map based on Longitude (generated) and Latitude (generated). Color shows sum of AIC Resids. Details are shown for State and County.

Residual Heat Map through Forward/Backward Selection



Map based on Longitude (generated) and Latitude (generated). Color shows sum of AIC Resids. Details are shown for State.

Conclusion

- **Use of these models:**
  - If you have current county information, you can predict income and unemployment levels
  - If you have a projection of where the county is going in the future, these models can determine what the unemployment and income levels may be
  - Look at variables to determine which conditions could be improved to increase income or lower unemployment

- **Future study:**
  - Refit these models when the 2020 census data comes out
  - Use these models to predict what income and unemployment may look like for the 2020 census