PROJECT II: PREDICTIVE ANALYTICS PROJECT

MaDonna Connors

Levkoff

18 November 2020

Executive Summary

The business relevant dataset featured previously was originally extracted from the Federal Reserve of St. Louis (FRED). The dataset decided on were factors affected or influencing the Great Recession by U.S. state in the year 2008. 2008 is particularly relevant to the overall goal of the analysis because the U.S. was in the middle of The Great Recession at this time. By definition, a recession is GDP, supported by several other factors. As several industries experienced a great collapse, many lost their jobs, homes, and thus individuals saw a large decrease in their personal income. Previously, we found that the most significant positive relationships discovered between variables were between the average resident population by state and the average unemployment rate, as well as between the real total GDP and average per capita personal income. There was also a negative relationship between average unemployment rate and average per capita personal income as well as the percentage of residents with a Bachelor's Degree or higher and the unemployment rate. We found that part of this could have been due to an economic downturn in 2008 that had a strong influence mainly on individuals in non-essential fields who usually earn a smaller amount of income.

The goal in this project is to use predictive modeling, or linear regression, to describe the relationships we found previously and decide which model fits the test data best. Average Unemployment Rate was chosen as the dependent variable (our variable of interest or focus) in the dataset because it fluctuates depending on other variables, such as the percentage of residents in a state's population or the percentage of residents with a Bachelor's Degree or higher, specifically during the recession. The variables represented after original data cleaning are State, Year, Average Unemployment Rate, Average Per Capita Personal Income, Real Total GDP, Average Resident Population, and Bachelor's Degree or Higher, as outlined with descriptions below.

STATE: This dataset is broken down by each U.S. state in alphabetical order. Although The Great Recession did affect other countries, the focus of this analysis is on the effects within the U.S.
YEAR: The year 2008 was chosen for this dataset because it helps to depict the middle of The Great Recession that has a large influence on each other variable that is chosen.
AVERAGE UNEMPLOYMENT RATE: This variable is identified as the percentage of the state's labor force that is unemployed at this given time.
AVERAGE PER CAPITA PERSONAL INCOME: This variable represents the average income that is earned per person in each given state from all sources in 2008 and is also represented as a dependent variable in the analysis because it is expected to rise or fall in response to other variables.
REAL TOTAL GDP: This variable represents the value of all goods and services produced by each state in 2008. Real GDP was chosen because it is adjusted for inflation.

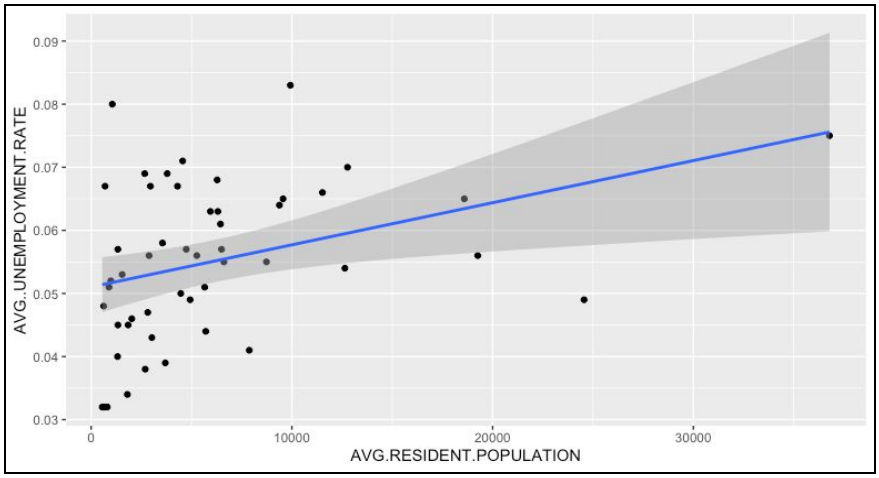
RESIDENT POPULATION: This variable represents the average number of persons (in thousands of persons) living in each given state during 2008.
BACHELOR'S DEGREE OR HIGHER: This variable is the percentage of each state's population that holds a Bachelor's Degree or higher and is represented as an independent variable in terms of the analysis.

In R Studio, the dataset was already clean but had to be broken up into training and test data in order to create the models.

```
Training <- df[train_ind, ] #pulling rows at random for training
Testing <- df[-train_ind, ] #pulling rows at random for testing

#checking dimensions for each partition - training/testing
dim(Training)
dim(Testing)
```

Diagnostic Results

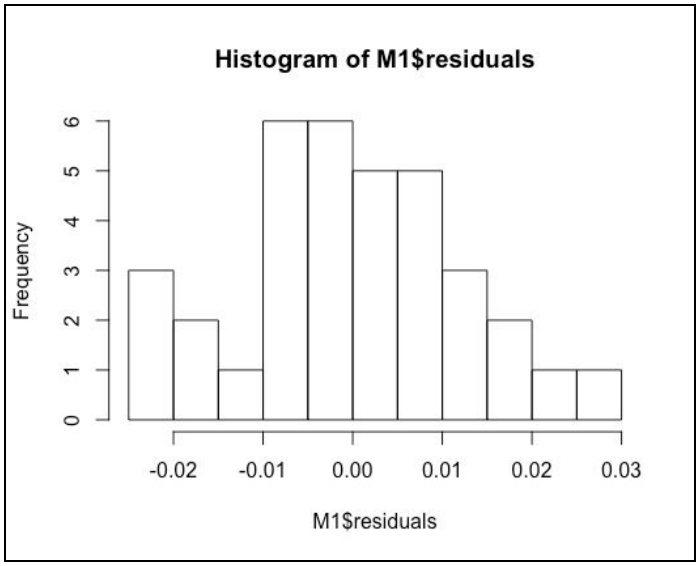
Model 1: AVG..UNEMPLOYMENT.RATE = B0 + B1*AVG.RESIDENT.POPULATION + u



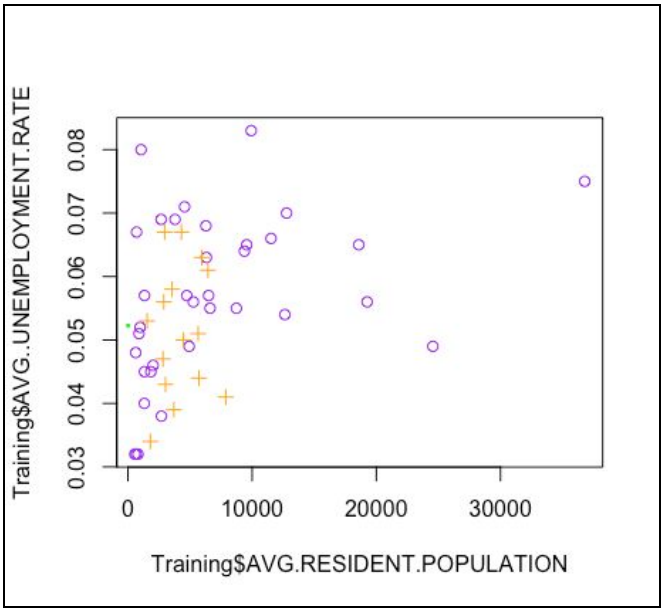
```
Call:
lm(formula = AVG..UNEMPLOYMENT.RATE ~ AVG.RESIDENT.POPULATION,
    data = Training)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0207583 -0.0078078 -0.0003591  0.0067603  0.0270839

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.225e-02  2.808e-03  18.609   <2e-16 ***
AVG.RESIDENT.POPULATION 6.282e-07  2.699e-07   2.328   0.0262 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01241 on 33 degrees of freedom
Multiple R-squared:  0.141,     Adjusted R-squared:  0.115
F-statistic: 5.418 on 1 and 33 DF,  p-value: 0.02621
```

First, Model 1 was built with the above regression equation. From returning the model coefficients and receiving the diagnostic output for this model, we can see that, for the intercept when we set x equal to 0, the predicted y variable is 5.225379e-02. For the slope, we can see that a 1 person increase in the average resident population increases the average unemployment rate by 6.281562e-07. Furthermore, we can see that for the average resident population there is an individual p-value of 0.0262 returned, which is less than the alpha level of 0.05. Thus individually this variable is statistically significant. This makes sense because as we found previously, the greater a state's population, the more likely they are to have a higher unemployment rate because of the density of residents.

For the f-statistic, the p-value returned is 0.02621, which is also less than the alpha level of 0.05, meaning the model is jointly significant and with 95% confidence, we can reject the null hypothesis and conclude that there is a relationship between at least one variable (average resident population) and unemployment rate. The Multiple R-squared is 0.141, meaning that about 14% of the variation in 'Avg...Unemployment.Rate' is explained by the variation in 'Avg.Resident.Population.' The Adjusted R-squared is 0.115, which will be used to compare with the coming models. At first glance at the histogram featured below, it appears that the residuals in this model are skewed, however with the Jarque Bera Test for normality, a p-value of 0.8302 is returned which is greater than the alpha value of 0.05, thus we can fail to reject the null hypothesis and conclude that the data is slightly normal. RMSE was used to compute the in-sample and out-of-sample error for each model. For Model 1 the in-sample error was 0.01204793, meaning on average we were off by this many units in the training data.




```
        Jarque Bera Test

data:  M1$residuals
X-squared = 0.37212, df = 2, p-value = 0.8302
```

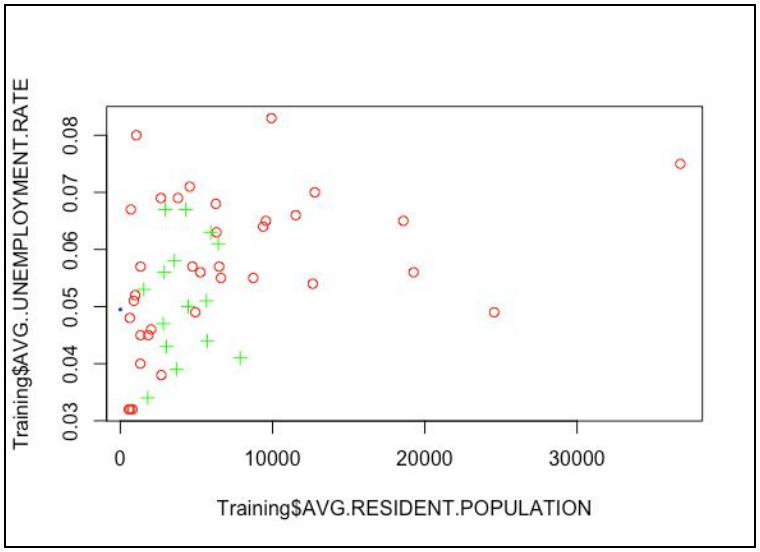


Model 2: AVG..UNEMPLOYMENT.RATE = B0 + B1*AVG.RESIDENT.POPULATION + B2*AVG.RESIDENT.POPULATION^2 + u

```
Call:
lm(formula = AVG..UNEMPLOYMENT.RATE ~ AVG.RESIDENT.POPULATION +
    AVG.RESIDENT.POPULATION2, data = Training)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0194574 -0.0070923 -0.0004669  0.0059641  0.0289885

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               4.949e-02  3.522e-03  14.052 3.08e-15 ***
AVG.RESIDENT.POPULATION   1.477e-06  7.146e-07   2.067   0.0469 *
AVG.RESIDENT.POPULATION2 -2.869e-11  2.240e-11  -1.281   0.2094
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01229 on 32 degrees of freedom
Multiple R-squared:  0.1829,    Adjusted R-squared:  0.1319
F-statistic: 3.582 on 2 and 32 DF,  p-value: 0.03946
```

For Model 2, a new variable, AVG.RESIDENT.POPULATION^2 Model 2 was created as a square of the old variable in the previous model and added to the dataset. From the diagnostic output above we can see that the individual ‘Avg.Resident.Population^2’ variable has a p-value of 0.2094, which is greater than the alpha level of 0.05, and is therefore not statistically significant, meaning we fail to reject the null hypothesis with 95% confidence and can conclude that there is no relationship between this variable in terms of explaining ‘Avg.Unemployment.Rate.’. In this model, only the original ‘Avg.Resident.Population’ variable is statistically significant at the 0.05 alpha level. However, when looking at the f-statistic, the p-value is 0.03946, which is less than than the alpha level of 0.03946, which confirms that at least 1 x variable in the model (‘Avg.Resident.Population’) is related to the y-variable statistically and the model variables are jointly significant together. In terms of residual normality, a p-value of 0.8302 is returned in the JB Test, which is greater than the alpha value of 0.05, thus we can fail to reject the null hypothesis and conclude that the data is normally distributed. In this model, the Multiple R-squared is 0.1829 meaning we can only explain about 18% of why unemployment rate fluctuates from state to state with the ‘Avg.Resident.Population’ variable. This is an increase from the previous model potentially meaning there is a better model fit. For Model 2 the in-sample error was 0.01175037, meaning on average we were off by this many units in the training data, which so far is means the model has a better fit than Model 1.

```
        Jarque Bera Test

data:  M2$residuals
X-squared = 0.70039, df = 2, p-value = 0.7046
```



Model 3: AVG..UNEMPLOYMENT.RATE = B0 + B1*AVG.RESIDENT.POPULATION + B2*AVG..PER.CAPITA.PERSONAL.INCOME + BACHELOR.S.DEGREE.OR.HIGHER + u

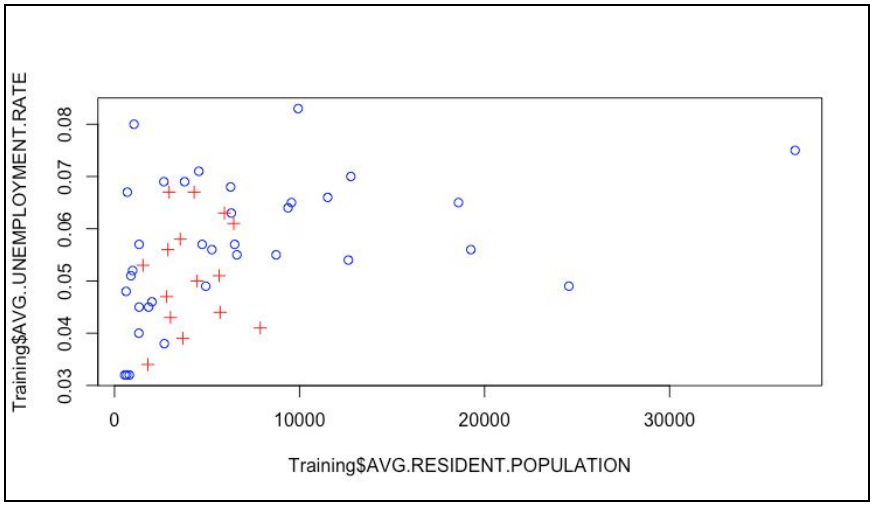
```
Call:
lm(formula = AVG..UNEMPLOYMENT.RATE ~ AVG.RESIDENT.POPULATION +
    AVG..PER.CAPITA.PERSONAL.INCOME + BACHELOR.S.DEGREE.OR.HIGHER,
    data = df)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0197496 -0.0064648  0.0001299  0.0064223  0.0292435

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    6.750e-02  1.109e-02   6.089 2.13e-07 ***
AVG.RESIDENT.POPULATION        7.121e-07  2.503e-07   2.846   0.0066 **
AVG..PER.CAPITA.PERSONAL.INCOME -4.621e-07  4.422e-07  -1.045   0.3015
BACHELOR.S.DEGREE.OR.HIGHER     5.172e-03  5.680e-02   0.091   0.9278
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01174 on 46 degrees of freedom
Multiple R-squared:  0.1736,    Adjusted R-squared:  0.1197
F-statistic: 3.221 on 3 and 46 DF,  p-value: 0.03115
```

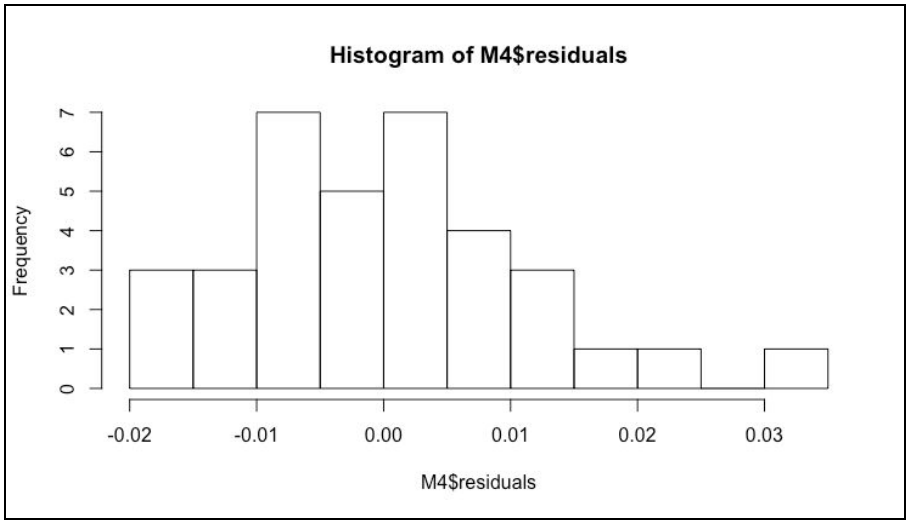
Model 3 is a model with multiple independent variables. From the diagnostic output above we can see that ‘Avg..Resident.Population’ is still the only statistically significant variable, with a p-value of 0.0066. ‘Avg..Per.Capita.Personal.Income’ (p-value = 0.3015) and ‘Bachelor’s.Degree.or.Higher’ (p-value = 0.9278) both have p-values that are greater than the alpha level of 0.05, thus we fail to reject the null hypothesis and conclude that there is no relationship between these independent variables and the dependent variable of ‘Avg..Unemployment.Rate.’ However, again the f-statistic p-value is 0.03115, which is less than

the alpha level 0.05, meaning at least one of the x variables in the model ('Avg.Resident.Population') must be related to the y variable statistically ('Avg..Unemployment.Rate'), thus the model is jointly significant. In this model, the Multiple R-Squared is 0.1736, meaning that about 17% of the variation in 'Avg.Unemployment.Rate' can be explained by the variation in the independent variables, and other variables in the dataset that were not included in this model account for the other 83%. The Adjusted R-squared value is 0.1197, which has decreased from the previous model because of the addition of more independent variables. For Model 3 the in-sample error was 0.1184575, meaning on average we were off by this many units in the training data. While the model improved from Model 1 to Model 2, it declined in fit from Model 2 to Model 3 in terms of the training data.

MODEL 4: AVG..UNEMPLOYMENT.RATE = B0 + B1*ln_AVG.RESIDENT.POPULATION + u

```
Call:
lm(formula = AVG..UNEMPLOYMENT.RATE ~ ln_AVG.RESIDENT.POPULATION,
    data = Training)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0180948 -0.0083825 -0.0005405  0.0059186  0.0304863

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.010640   0.013882   0.766  0.44885
ln_AVG.RESIDENT.POPULATION 0.005585   0.001670   3.344  0.00207 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01157 on 33 degrees of freedom
Multiple R-squared:  0.2531,    Adjusted R-squared:  0.2305
F-statistic: 11.18 on 1 and 33 DF,  p-value: 0.002067
```






```
        Jarque Bera Test

data:  M4$residuals
X-squared = 1.8748, df = 2, p-value = 0.3917
```

Our final model is a logarithmic model built from the training data. From the diagnostic output, we can see that 'Ln_Avg.Resident.Population' is statistically significant (p-value 0.00207) and with 95% confidence, we can reject the null hypothesis and conclude that there is a relationship between this independent variable and the 'Avg.Unemployment.Rate' y variable. The f-statistic p-value is 0.002067, which is less than the alpha level of 0.05, thus the model is jointly significant and we reject the null hypothesis. Therefore 'Ln_Avg.Resident.Population' is related to the dependent variable statistically. The Multiple R-squared is 0.2531, meaning that we can explain about 25% of why 'Avg.Unemployment.Rate' fluctuates from state to state with the 'Ln_Avg.Resident.Population' variable and all other variables in the dataset account for the other 75%. A p-value of 0.3917 is returned in the JB Test, which is greater than the alpha value of 0.05, thus we can fail to reject the null hypothesis and conclude that the data is normally distributed. The Adjusted R-squared is now 0.2305, an increase from the previous model meaning the new variable has improved the model fit. For the final model, Model 4, the in-sample error was 0.0112344, meaning on average we were off by this many units in the training data. This is the lowest sampling error out of each model, and therefore Model 4 performs the best on the training data. Model 4 also has the highest Multiple R-squared, which means it explains more of this dataset. In order to figure out which model performs the best on the testing data, we computed the out-of-sample error in the next section.

```
> #COMPARISON OF IN-SAMPLE MODEL PERFORMANCE BY RMSE
> RMSE_1_IN #MODEL WITH ONLY LINEAR TERM
[1] 0.01204793
> RMSE_2_IN #MODEL WITH LINEAR AND QUADRATIC TERM
[1] 0.01175037
> RMSE_3_IN #MODEL WITH MULTIPLE INDEPENDENT VARS.
[1] 0.01184575
> RMSE_4_IN #LOGARITHMIC MODEL
[1] 0.0112344
```

Predictions

```
> #COMPARISON OF OUT-OF-SAMPLE MODEL PERFORMANCE BY RMSE
> RMSE_1_OUT #MODEL WITH ONLY LINEAR TERM
[1] 0.01046325
> RMSE_2_OUT #MODEL WITH LINEAR AND QUADRATIC TERM
[1] 0.01056546
> RMSE_3_OUT #MODEL WITH MULT. INDEPENDENT VARS.
[1] 0.009749984
> RMSE_4_OUT #LOGARITHMIC MODEL
[1] 0.01111436
```

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

We used the Root Mean Square Error (RMSE) metric to benchmark each model against one another to see how accurate each model is in terms of prediction. Here we are predicting based on a new dataset (i.e. testing data) without memoization. In order to come to this conclusion, we calculated the out-of-sample error for each model. For Model 1, the out-of-sample error was 0.01046325, which was actually lower than the in-sample error for the same model. For Model 2, the out-of-sample error was 0.01056546, which is also lower than the in-sample error of the same model. For Model 3, the out-of-sample error was 0.009749984, which is also lower than the in-sample error of the same model and is the lowest out-of-sample error out of every model. Lastly, Model 4's out-of-sample error is 0.01111436 which again is slightly lower than the in-sample error of the same model. Model 4 is the best model in terms of the training data and for the best fit, but Model 3 has the best performance in terms of the test data, meaning it predicts the best on new observations, which is critical to success in predictive analysis.

Conclusion

Ultimately, 4 different models were built to see which model performs the best in terms of predictive analytics. These models were a Model 1 with a linear term, Model 2 with a linear and quadratic term, Model 3 with multiple independent variables, and Model 4 which was a logarithmic model. Although the only statistically significant variable in the dataset is 'Avg.Resident.Population,' each model was jointly significant meaning at least the one variable must be related to the y-variable, which was 'Avg.Unemployment.Rate.' 'Avg.Resident.Population' still appears to have a strong positive correlation with 'Avg.Unemployment.Rate.' In conclusion, although Model 4 (logarithmic model) was the best model in terms of fit on the training data, Model 3 (multiple linear regression) is the recommended model because it has the lowest out-of-sample error of any model, meaning it predicts the relationships with 'Avg.Unemployment.Rate' the best on the training data set of previously hidden observations.