Predictive Analytics Project

Victoria Roberts

Dr. Levkoff

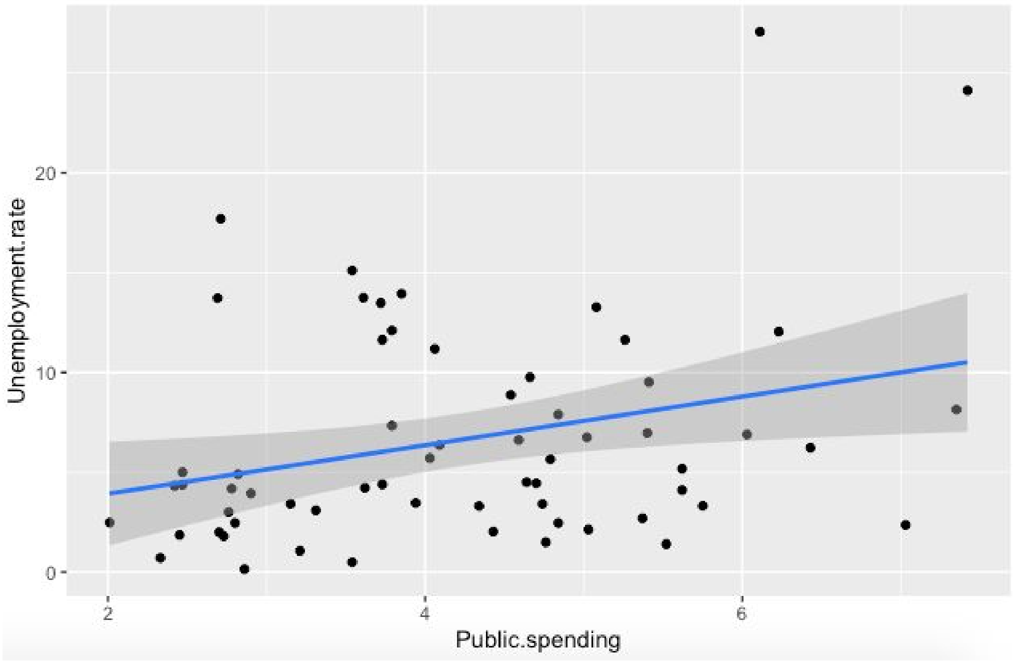
ECON 494 (1:00 pm)

November 18, 2020

Executive Summary

The data collected is about unemployment across countries and the variables that might cause or correlate the effects of unemployment. The dataset is focused particularly on 2017 and the fluctuations there might have been in the unemployment rate in that year. This cross-sectional dataset helps understand the relationship between unemployment and other variables. Variables such as: Inflation Rate, Population, Public Spending on Education, and Nominal GDP. The data was sourced from publicly available information on gapminder and the global economy. The variables were chosen because of their relevance to each other. The tradeoff between inflation rate and unemployment rate have been the most monitored economic indicators. These two variables have an inverse correlation, this means that while one increases the other decreases. Next, given the size of a country's population could help analyze the unemployment rate. An educational variable was incorporated into the dataset to be able to analyze its relationship with unemployment. The education a person has received or the availability of education can help understand unemployment. Lastly, changes in GDP can substantially burden the unemployment rate. To identify relationships between variables, I used histograms and scatter plots with a smoothed geom. Smoothers help convey the degree of association between variables to make a more compelling visualization. Shown below is a scatter plot between unemployment and public spending on education. Based on this, there is a slight positive correlation between these two variables. This will help gain further insight to analyze and predict the fluctuations of the unemployment rate.

Figure 1: Scatter Plot Between Unemployment Rate and Public Spending on Education



Proposal of Models

Before proposing the different models that will be used to test unemployment rate, the data was partitioned into two distinct groups: the training data and the testing data. Typically, 70% of the dataset is used to train and build a model and the other 30% is used to test or benchmark a model's performance. Once the data is partitioned, we can start building and estimating models to help predict the y-variable. Four different linear regression models will be involved in predicting unemployment rate. The first model that will be used is a simple linear regression model with one x-variable to test the likelihood of the variable predicting y. Next, a quadratic linear regression model will be used to test the same x-variable plus its quadratic form in predicting the y-variable. The logarithmic regression model will be built using the training data to test the relationship between the x and y-variables. The logarithmic regression takes the log of the x value and tests to see if there is a statistical significance. Lastly, the fifth model that will be used is the multiple linear regression. This particular model tests the y-variable against more than one regressor on the right side of the tilda.

Results

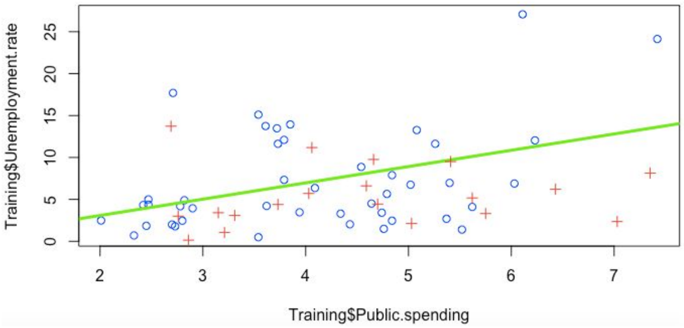
Model 1: Linear Regression Model

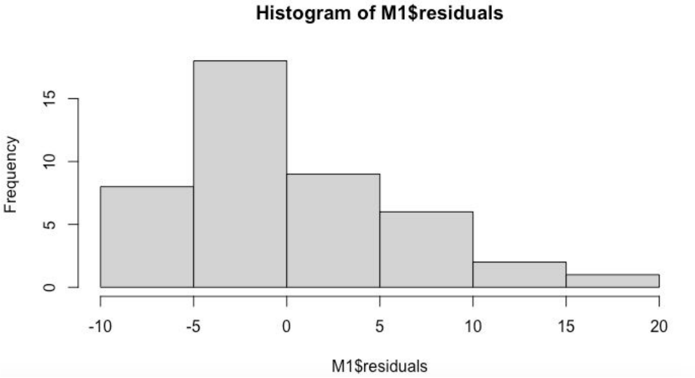
From the training data partition, a linear regression model was conducted to test the likelihood between public spending on education and unemployment rate. The model will help describe if the x-variable, public spending on education, predicts the y-variable, unemployment rate. In the coefficients table of Model 1, spending on education is statistically significant at the 1% level. Surprisingly, the intercept did not show any significance (**). The intercept is a constant variable and is intrinsically not significant or important given that it is not a regressor but the variable we are trying to predict. Furthermore, the model is jointly significant with an F-test of 8.375 (p-value= .006). The p-value is less than the alpha of .05, therefore the model is statistically significant. With a 95% confidence level, we can reject the null hypothesis that assumes all betas are zero except for the slope. The multiple r-squared shows how much of the variation in y is explained by the variation in x. In other words, spending on education explained about 16% of the fluctuation in the unemployment rate. Therefore, 84% of other variables must account for the rest of the variation. The residual error shows the errors on each of the data points relative to that of the regression line. The regression equation for this specific model is:

- *Unemployment Rate= B0 + B1*Public Spending on Education +u*

In addition, the model has a residual standard error of 5.601. This means that the model underpredicted the values between the training and testing data. The actual values are well above the predicted values on the regression line. This is shown in the visualization below, portraying the data and the regression plot overlaid. Furthermore, a histogram was created to check the normality of the residuals. Based on the histogram below, the data is skewed to the right portraying that the residuals are not normally distributed. To further test for normality, a Jarque Bera Test was conducted and showed a p-value of .01.

Figure 2&3: Scatter plot between the training and testing data and histogram to check normality





Model 2: Quadratic Regression

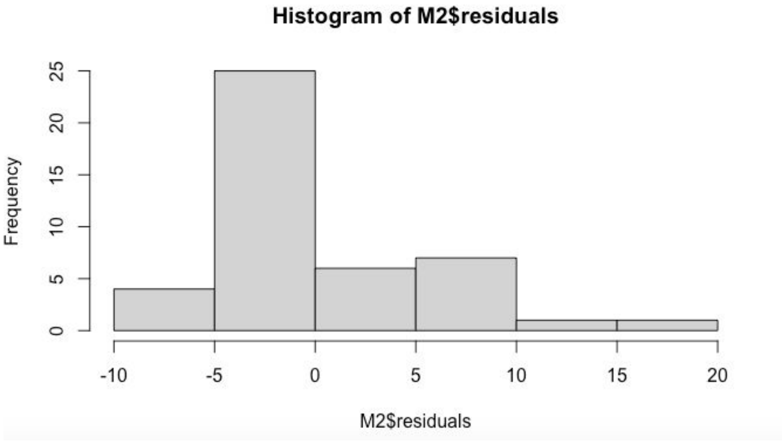
Given that the regression line was underpredicting the data and was not fitting the line into the data points, a second model was built with an additional variable. Model 2 was implemented to try to improve the overall fit of the data. All models will be built with the same training data and will be used to benchmark each model's performance. The regression line was not as accurate in predicting the points in the data, therefore instead of using a line to predict, a quadratic regression was incorporated which is an extension of the linear regression model. Model 2 uses the same variables as Model 1 but an additional variable is added, the quadratic form of the x-variable. Specifically, the coefficients table shows that the normal variable of x is not statistically significant while the squared value is statistically significant at the 5% level. Although not all variables were statistically significant, the model as a whole was jointly significant. This means that Model 2 can reject the null hypothesis and conclude that at least one of the variables is helping to predict the y-variable (p-value- .001).

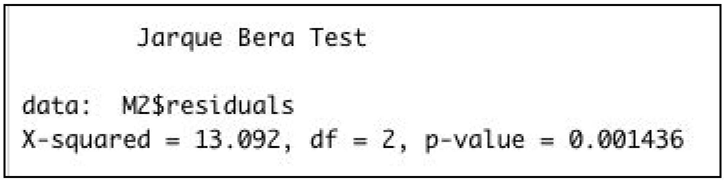
When adding the quadratic variable, the model's multiple r-squared increased from the first model. Specifically, 27% of the variation in the y-variable is explained by the data. The multiple r-square tends to increase the more variables that are added because the model is fitting the data better and is explaining the variation between variables. The residual standard error decreased from the first model to the second, meaning that there is less error between the data points and the regression line. Although the residual error did not change much, the quadratic line still fits the data better than Model 1. The quadratic equation used to build the model 2 was:

- *Unemployment Rate= B0 + B1*Public Spending on Education + B2*Public Spending on Education^2 +u*

Based on the in sample and out sample predictions that were generated, each showed a value of 5.086 and 6.622 respectively. The in sample prediction is generated with the training data and the out of sample predictions are generated with the testing data. In Model 1, the in sample data was 5.472 and the out of sample data prediction was 4.985. What does this mean? This means that from Model 2 in sample data is better fitting the dataset, while the Model 1 out of sample is a better predictor of the dataset. Next, the JB test was conducted to test for normality in the residuals of Model 2. The histogram shows less skewness compared to Model 1, but is still slightly skewed to the right. Figure 4 portrays a skewed distribution towards the right, signifying that the residuals are not normally distributed. On the right of the histogram, the JB Test hypothesizes a test for the residuals of the model. Given that the p-value is less than .05, we can conclude to reject the null hypothesis stating that the data is normal.

Figure 4&5: Histogram of residuals and JB Test of Normality



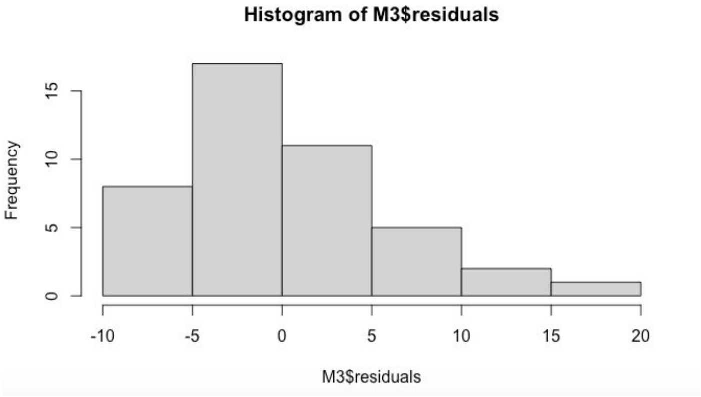


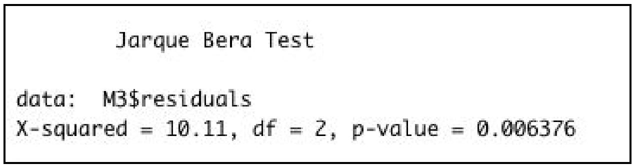
```
        Jarque Bera Test

data:  M2$residuals
X-squared = 13.092, df = 2, p-value = 0.001436
```

Model 3: Logarithmic Regression

To further analyze a different class of regression model, Model 3 is conducted by using the logarithmic regression of the x-variable. The equation consists of the y and the x variable, but instead of the normal x-variable the log of the variable x is used to conduct the test. Specifically, the x-variable is statistically significant at the 5% level. In addition, the model as a whole is also statistically significant with an F-statistic of 6.94, given that the p-value is less than the alpha level of significance (p-value=.011). Surprisingly, the multiple r-squared decreased to 14% of explained variation. This means the Model 2 is better in fitting and predicting the model to the data than the logarithmic regression. On average we are about 5.683 off when we try to predict the unemployment rate based on the x-variable. When testing for normality, the histogram portrays the same skewed distribution as Model 1. Given the JB Test, with a p-value less than the alpha level, we can conclude to reject the null hypothesis. The more complexity in the model, the higher the multiple r-squared, this would make sense given that Model 2 has a higher multiple r-squared than that of Model 3 with only one x-variable.

Figure 6&7: Testing for normality and distribution with a Histogram and JB Test





```
        Jarque Bera Test

data:  M3$residuals
X-squared = 10.11, df = 2, p-value = 0.006376
```

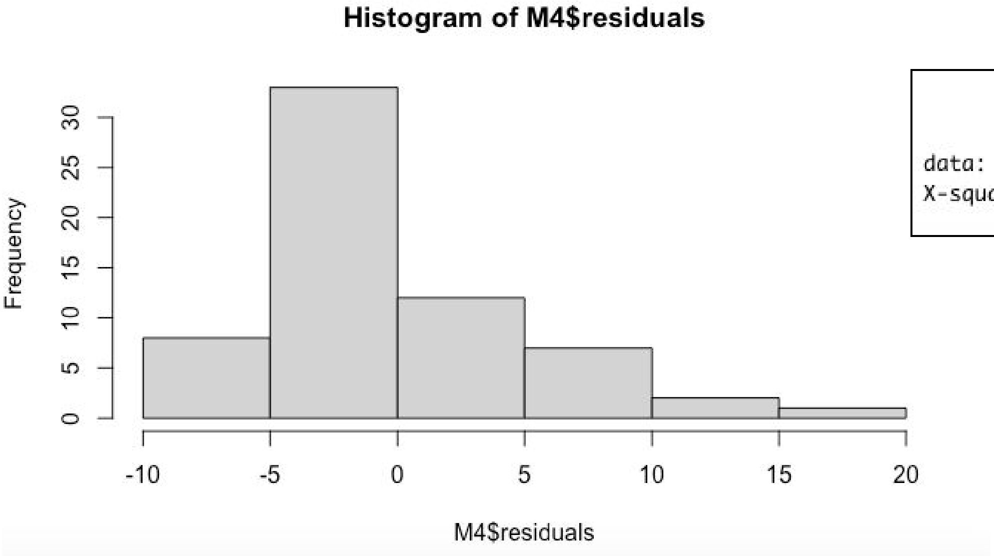
Model 4: Multiple Linear Regression

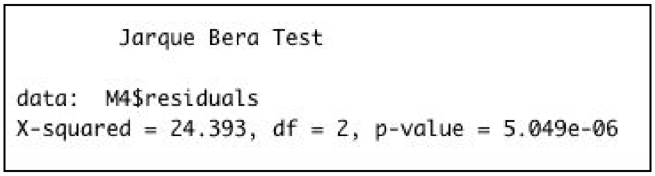
Lastly, Model 4 is a model with multiple independent variables testing the likelihood that they will help predict the y-variable. The multiple regression equation is written as follows:

- *Unemployment.rate= B0 +B1*Public.spending + B2*GDP + B3*Inflation.rate + B4*Population + u*

The more x-variables are added to the regression equation the more complex it becomes when referring to the r-squared. Nevertheless, it is a balanced equation, meaning that the more x-variables that are added may fluctuate the multiple and adjusted r-squares. Based on the regression conducted, it was evident that variables such as GDP, Inflation rate, and Population were not statistically significant. These regressors have no relationship and do not help predict the variation in unemployment rate. Interestingly, inflation rate and population are negative coefficients. This means that they negatively impact the rate of unemployment. In addition, public spending on education is significant in determining the effect of unemployment at the 5% level (p-value=.02). Public spending on education has a t-test of 2.304, indicating its relationship between the y-variable. The test, as a whole, is not jointly significant given its p-value of 0.182. In addition, the multiple r-squared decreased from Model 3 to Model 4. The variables explain 10% of the variation in the y-variable. This is very low and could mean that the variables that were chosen do not correlate and help predict the y-variable. The adjusted r-squared tends to be lower than the multiple r-squared because it is restrictive to the addition of x-variables and therefore gives a better picture of the r-squared. When testing for normality, the p-value was very small which indicates that the data is not normally distributed. The histogram shows the average skewness of the other models.

Figure 8&9: Histogram to test distribution of residuals and JB Test to test normality





```
        Jarque Bera Test

data:  M4$residuals
X-squared = 24.393, df = 2, p-value = 5.049e-06
```

Predictions

By using the testing data partition, the data will be used to benchmark the models and evaluate their performance. We test using the testing data because it is data that we haven't seen, or memorized. In other words, it is the data that we have not used to build the models so it will show a better picture and the better model. The purpose of these models is to learn and predict what might cause the fluctuations in the rate of unemployment. Using out of sample performance is key to determine the true estimate between models and their respective predictions. The error metric that was used was RMSE to benchmark the different models against each other. RMSE measures the average deviation in the actual data relative to the prediction and contains the same units of measurement as the output variable, y. The best in sample prediction is Model 2 with an RMSE of 5.08. The ranking of the in sample models go as follows: M2(5.08)>M4(5.46 not sig.)>M1(5.47)>M3(5.55). This states that Model 2 is the best model that fits the data correctly. In sample data utilizes the training dataset which is used throughout building the models. Furthermore, the best model that predicts the out of sample error is Model 4 with an RMSE of 4.32. Model 4 was not jointly significant, therefore I believe it is not the best predictor of the y-variable. If Model 4 were to have been statistically significant, then this model would be the best in estimating the variation in unemployment rate. Based on this data, Model 3 would be the best model to characterize and predict the y-variable. The ranking of RMSE for out of sample data is M4(4.32 not sig.)>M3(4.67)>M4(4.98)>M2(6.62). For this particular project, out of sample data is the best way to compare the models performance with data that has not been memorized.

Conclusion:

Based on the analysis and model comparison, Model 3 is the best indicator of the y-variable, unemployment rate. The goal of this project is to predict, with the best model, the rate of unemployment. Given that Model 4 is not jointly significant, I cannot recommend this model to truly test and predict y. If the model were to have been significant then this would have been the best prediction model. Nevertheless, I propose Model 3 is the best model to predict the variable of interest. Having a lower RMSE signifies that there is less error in the performance of the model. The less error in a model leads to a more specific and accurate representation of the x variables affecting y.