494 ECO PROJ 2

DATA ANALYSIS ON FACTORS AFFECTING INFLATION

University of San Diego

Dr. Levkoff

November 18, 2020

# Executive Summary

This project is designed to build linear regression model in order to find relationships of

inflation and other factors. The data set is searched and downloaded from the Gapminder website. All of the variables are formed by a “mix-and-match format”, with each variable downloaded separately and combined afterwards. There are 7 variables in the data set, including annual inflation, GDP per capita growth, merchandise trade, income per person, Gini coefficient, annual population growth, and developed country. In order to work with the data better, the names of the variables are changed in a simple way and the class of data are change into integer if needed. All of the above processes are shown in the code from line 6-36.

Based on Project 1, there is a relationship between inflation and factors, such as income, GDP and population growth. Also based on the previous project, whether a country is developing country or developed country also takes into account in finding the relationship; and being as a developing country or developed country does have different impact on inflation. Furthermore, all relationships can be identified toward a “flat” relationship instead of a “steep” relationship, which indicates that those relationships are not too strong.

Graphical user interface, text, application

Description automatically generated

Figure 1 R Script on Basic Data Information

In order to graph and to do the JB test for normality, 3 libraries are loaded to R code (line 45-47). And two nonlinear transformation are created in order to build nonlinear regression models for relationship between inflation and income (line 51-52). And in order to build the regression model, data set has been divided into training and testing partitions, with the fraction of sample data as 70%. To fix the training and testing data, a set of seed is coded in line 65 in Figure 2. As a result, the dimension or observations for training data is 122, which is 70% of all observations, and the dimension for testing data is 53. (Figure 3)

Graphical user interface, text, application, email

Description automatically generated

Figure 2 R Script on Data Partition

A picture containing company name

Description automatically generated

Figure 3 Dimensionality for Training and Testing Data

1. Linear Regression Models

There are 6 models build in this project. 4 of the models are linear models, the rest 2 are

nonlinear models. For each mode, the dependent variable is inflation because the models are used to predict the relationships between inflation and other factors. The first model is a linear model between inflation and income, which is the biggest factor that influences inflation on project 1 (shown in Figure 4). The training data is used to build the model, and some indicators are showing whether this is a good model for the data. According to the summary in Figure 5, income has a negative relationship with inflation; in other words, as income goes up by 1 unit, inflation goes down by 5.62\*10^-7. The p-values for both the variable and the model is small enough, which means the model is significant. The R squared states that about 8.84% of the data are explained by this model, which is not enough to be a good fit model for our data set. Based on the histogram in Figure 6 and the Jarque Bera Test in Figure 5, the p-value shows the null hypothesis is rejected, which means that the residuals cannot be concluded as normally distribution. Figure 7 shows a graph in 2D against both data partitions. And the in sample error and out sample error will be compared later in part 3 and 4 of this project.

Graphical user interface, text, application, email

Description automatically generated

Figure 4 R Script for Model 1

Text, letter

Description automatically generated

Figure 5 Summary and Normality Test for First Model

Chart, scatter chart

Description automatically generated

Figure 6 Graphs for Relationship between Inflation and Income

Chart, scatter chart

Description automatically generated

Figure 7 Graph in 2D against Both Data Partitions

Model 2 is a linear model with inflation as the dependent variable and income and GDP as the independent variables. According to summary in Figure 9, income has a negative relationship to inflation, while GDP has a positive relationship to inflation. The p-values for income and the model is small enough, which means the model is significant. But the p-value for GDP is too big to be significant. The R square is 9.96%; although it is higher than model 1, it is not enough to be a good fit model for our data set. Based on the histogram and the Jarque Bera Test in Figure 9, the p-value shows the null hypothesis is rejected, which means that the residuals cannot be concluded as normally distribution. The in sample error and out sample error will be compared later in part 3 and 4 of this project.

Text, letter

Description automatically generated

Figure 8 R Script for Model 2

Text

Description automatically generated

Figure 9 Summary and Normality Test for Model 2

Chart, histogram

Description automatically generated

Figure 10 Histogram for Model 2

Model 3 is a linear model with inflation as the dependent variable and income, GDP and trade as the independent variables. According to summary in Figure 12, income has a negative relationship to inflation, while GDP and trade have positive relationships to inflation. The p-values for income and the model is small enough, which means the model is significant. But the p-values for GDP and trade are too big to be significant. The R square is 10.23%; although it is higher than model 1 and 2, it is not enough to be a good fit model for our data set. Based on the histogram and the Jarque Bera Test in Figure 12, the p-value shows that the null hypothesis is rejected, which means that the residuals cannot be concluded as normally distributed. The in sample error and out sample error will be compared later in part 3 and 4 of this project.

Text, letter

Description automatically generated

Figure 11 R Script for Model 3

Text

Description automatically generated

Figure 12 Summary and Normality Test for Model 3

Chart, histogram

Description automatically generated

Figure 13 Histogram for Model 3

Model 4 is a linear model with inflation as the dependent variable and income, GDP and population growth as the independent variables. According to summary in Figure 15, income and population growth have negative relationships to inflation, while GDP has a positive relationships to inflation. The p-values for income and the model is small enough, which means the model is significant. But the p-values for GDP and population growth are too big to be significant. The R square is 11.32%; although it is higher than previous models, it is not enough to be a good fit model for our data set. Based on the histogram and the Jarque Bera Test in Figure 15, the p-value shows the null hypothesis is rejected, which means that the residuals cannot be concluded as normally distribution. The in sample error and out sample error will be compared later in part 3 and 4 of this project.

Text, letter

Description automatically generated

Figure 14 R Script for Model 4

Text

Description automatically generated

Figure 15 Summary and Normality Test for Model 4

Chart, histogram

Description automatically generated

Figure 16 Histogram for Model 4

Model 5 is a model with inflation as the dependent variable and income, income^2 as the independent variables. According to summary in Figure 18, income and income^2 have negative relationships to inflation. The p-values for both are too big to be significant. The R square is 9.29%, which is not enough to be a good fit model for our data set. Based on the histogram and the Jarque Bera Test in Figure 18, the p-value shows the null hypothesis is rejected, which means that the residuals cannot be concluded as normally distribution. The in sample error and out sample error will be compared later in part 3 and 4 of this project.

Text

Description automatically generated

Figure 17 R Script for Model 5

Text

Description automatically generated

Figure 18 Summary and Normality Test for Model 5

Chart, scatter chart

Description automatically generated

Figure 19 Histogram and 2D Graph for partition for Model 5

Model 6 is a model with inflation as the dependent variable and log of income as the independent variables. According to summary in Figure 21, log of income have negative relationships to inflation. With one unit increase in log of income, the inflation decrease by 0.0058. The p-values for both are small enough to be significant. The R square is 3.25%, it is not enough to be a good fit model for our data set. Based on the histogram and the Jarque Bera Test in Figure 21, the p-value shows the null hypothesis is rejected, which means that the residuals cannot be concluded as normally distribution. The in sample error and out sample error will be compared later in part 3 and 4 of this project.

Graphical user interface, text

Description automatically generated

Figure 20 R Script for Model 6

Text, letter

Description automatically generated

Figure 21 Summary and Normality Test for Model 6

Chart, scatter chart

Description automatically generated

Figure 22 Histogram and Relationship Graph for Model 6

Chart, scatter chart

Description automatically generated

Figure 23 2D Graph for partition for Model 6

1. Diagnostic Results

Although the models show the relationships between inflation and other factors, there is not a model that fits the data well. But the relationships between income, GDP, trade and population growth are expected based on the previous project. Figure 24 shows the 6 equations for all models. According to the model, income, income^2, log income and population growth have negative relationships to inflation; GDP and trade have positive relationships for inflation. According to Figure 25, the in-sample error are shown. Among those in-sample error, model 4 has the lowest in-sample error (except for the 0 in-sample error for model 5), which shows that model 4 is the model that fits the data best.

A picture containing text

Description automatically generated

Figure 24 6 Model Equations

Text

Description automatically generated

Figure 25 In-Sample Errors for All 6 Models

1. Model Predictions

With the test data partition, the RMSE are calculated for each model. The out-of-sample

errors are calculated for all 6 models in Figure 26. According to the out-of-sample errors, model 1 has the lowest number for out-of-sample error, which shows that model 1 is the most against testing model. As mentioned above, model 5 has the lowest in-sample error, while model 1 has the lowest out-of-sample error. This is because model 5 does a better job in explaining the data, while model 1 does a better job in forecasting the model.

Text

Description automatically generated

Figure 26 Out-of-Sample Errors for All 6 Models

1. Conclusion

This project proves the relationships between inflation and other factors that were

found in the previous project. To summarize, income, income^2, log of income and population growth have negative relationships to inflation; GDP and trade have positive relationships for inflation. Among the 6 models, although none of the models fits the data well, nor does the models forecast the data well, model 5 has the smallest in-sample error and model 1 has the smallest out-of-sample error. In other words, mode 5 fits the data best and model 1 against the testing model the most. Since the latter one is what are expected, so model 1 are picked as the best model. However, there may be errors in collecting data that cause some outliers. Only with more observations and mode data set could we conclude the more accurate relationships between inflation and other factors. So there is need to involve more data from different countries and time period in order to further analyze the relationship between inflation and other factors.