MULTIPLE REGRESSION MODELLING

Group 7 Final Project

Talita Teixeira Santos & Ananyaa Shahi

**INTRODUCTION**

GDP stands for Gross Domestic Product. It measures the amount of goods and services produced within a country. It is a monetary value. This value is typically a good indicator of the overall economic health of a country. A country usually wants a higher GDP because that means they have a lot of resources, are sustainable, independent and can export goods and services. For this project, the GDP values of the United States were chosen to be looked at closer. The data set includes data from 1970 to 2023. The purpose of this project is to understand what factors affect GDP and might be used to predict GDP by finding the most accurate multiple regression equation. The equation includes variables and shows their magnitude, explaining how they affect GDP.

**DATASET DESCRIPTION**

There were 6 explanatory variables that were initially included in the model. The variables include unemployment rate, government expenditure, interest rate, export amount, import amount and whether the data was collected after the great recession of 2007-2009. These variables were all chosen because they are also indicators of how well a country is doing economically and are linked to goods and services production of the US. The unemployment rate and interest rate values are in percentage form. Unemployment rate is related to GDP because GDP is about production and unemployment rate measures the amount of people that are not working, therefore not producing goods or services. Government expenditure, export and import amount are in billions of dollars. These variables could be related to GDP because in some form, they indicate whether the economy is doing well or not. Government expenditure measures how much the country is spending. Export amount is the amount of goods and services being sold to other countries and import is the amount being bought from other countries. Whether or not the data was collected before the great recession of 2007-2009 is a qualitative variable. To be able to include this in the project, it needed to be assigned an indicator value. Since the recession is 3 years long, 2009 is used as the reference year. 0 is used for no and 1 is yes. All values until 2009 are 0 and 2010-2023 is 1. This will indicate if the recession has any effect on GDP. All the data was collected from the website FRED.stluisfed.org.

Equation 1:

The equation above is the initial equation for the project, prior to running any modelling. Y is GDP, is the intercept, is the coefficient to its respective variable and the slope of the equation.

**EXPECTED REALTIONSHIP OF PREDICTORS WITH PREDICTED(GDP)**

Below is the list of predictors and our expectation as to how it could be related to GDP:

|  |  |
| --- | --- |
| Unemployment Rate | The expectation is that unemployment rate and GDP have a negative relationship. |
| Government Expenditure | The expectation is that government expenditure and GDP have a positive relationship. |
| Import amount | Import amount is expected to have a negative relationship with GDP. |
| Export amount | Export amount is expected to have a positive relationship with GDP. |
| Interest Rate | It was expected that interest rate and GDP have a positive relationship. |
| Recession | The expectation is that post-recession periods would have a positive relationship with GDP. |

**STEPWISE REGRESSION MODELLING**

The forward-stepwise method was used to find the best multiple regression equation to predict GDP. After formulating the equation above and gathering data, step one consisted of building a correlation table for the variables. From there, the one with the highest correlation value to GDP was chosen to perform a single regression, which was import amount. Since the p-value was statistically significant, less than 5%, step 2 was carried out. Step 2 consisted of running regressions and adding a second variable one at a time to see how it affected the model. So, the import amount was kept constant and only one X variable was added to each regression. From there, the p-value was evaluated again for each regression. Step two dropped export amount, recession, interest rate and unemployment, only keeping government expenditure. So, for step three the X variables import amount and government expenditure were kept constant as the third X variable was again slowly added to the model one by one. After evaluating the p-values and adjusted R-Squared, the unemployment rate was chosen to move on to step 4. So now the consistent X variables are Import Amount, government expenditure and unemployment rate. Step 4 was the last step because all the X variables were insignificant as they were added one by one to the model.

**Equation 2:**

The equation above, equation 2, is the result equation after performing the stepwise method. The interest rate, recession and export variables were dropped and left with unemployment rate, government expenditure and import.

**RESULTS PER STEPWISE REGRESSION**

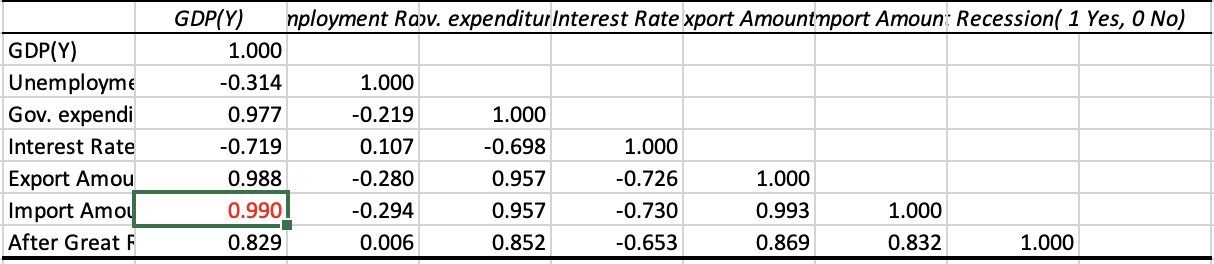
Below is the results we got about the predictors and their relationship to GDP:

|  |  |
| --- | --- |
| Import Amount | Contrary to our thought, the Import Amount has a positive effect on GDP. Per every billion dollars of amount imported, GDP increases by 3.67 billion dollars times; keeping Government Expenditure and Unemployment rate constant. |
| Government Expenditure | Government Expenditure aligns with our thoughts of having a positive effect on GDP. For every billion dollars of government spending, GDP increases by 1.59 billion dollars, keeping the unemployment rate and import amount constant. |
| Unemployment | As we had thought, Unemployment has an inverse relation with GDP. Per our regression equation, for every percent increase in unemployment, the GDP falls by 230.95 billion USD, keeping Government Expenditure and Import Amount constant. |

**APPENDIX**

**Regression Model:**

* We started the modelling by first calculating the correlation amongst the various predictors (Unemployment Rate %, Gov. expenditure, Interest Rate, Export Amount, Import Amount, After Great Recession- Yes or No) and the predicted variable, i.e., GDP. We found *Import Amount* as the predictor with the highest correlation (0.99) with GDP.



* In the first step, we ran the regression between *GDP* and *Import Amount.* We got almost zero for p-value (statistically significant at α = 5%) and 0.9795 for Adjusted R-Square. This suggested that the *Import Amount* is a good predictor for GDP, and it became the first predictor of our model.

A screenshot of a spreadsheet

Description automatically generated

* In Step 2, we included the rest of the variables (Unemployment Rate, Government Expenditure, Interest Rate, Export Amount, and After Great Recession- Yes or No) and ran regression with *GDP and Import Amount* one by one. Amongst all the predictors, we found *Government Expenditure* to be thebest one as the p-value (almost zero) was quite low and statistically significant at α=5%. Also, given that the adjusted R Square value was 0.9902, which is quite high, we decided to include *Government Expenditure* as our second predictor. Below is the summary of our findings:

A table with numbers and red text

Description automatically generated

* In Step 3, we ran regression tables for Unemployment Rate, Interest Rate, Export Amount, and After Great Recession-Yes or No with *GDP vs Import Amount & Government Expenditure. Unemployment Rate* turned out to be the best one amongst the rest of the predictors as it gave an Adjusted R-Square Value of 0.9925 and a quite low p-value (0.0002). Thus, the *Unemployment Rate* became our third predictor for the model. Below is the summary of our findings:

A screenshot of a spreadsheet

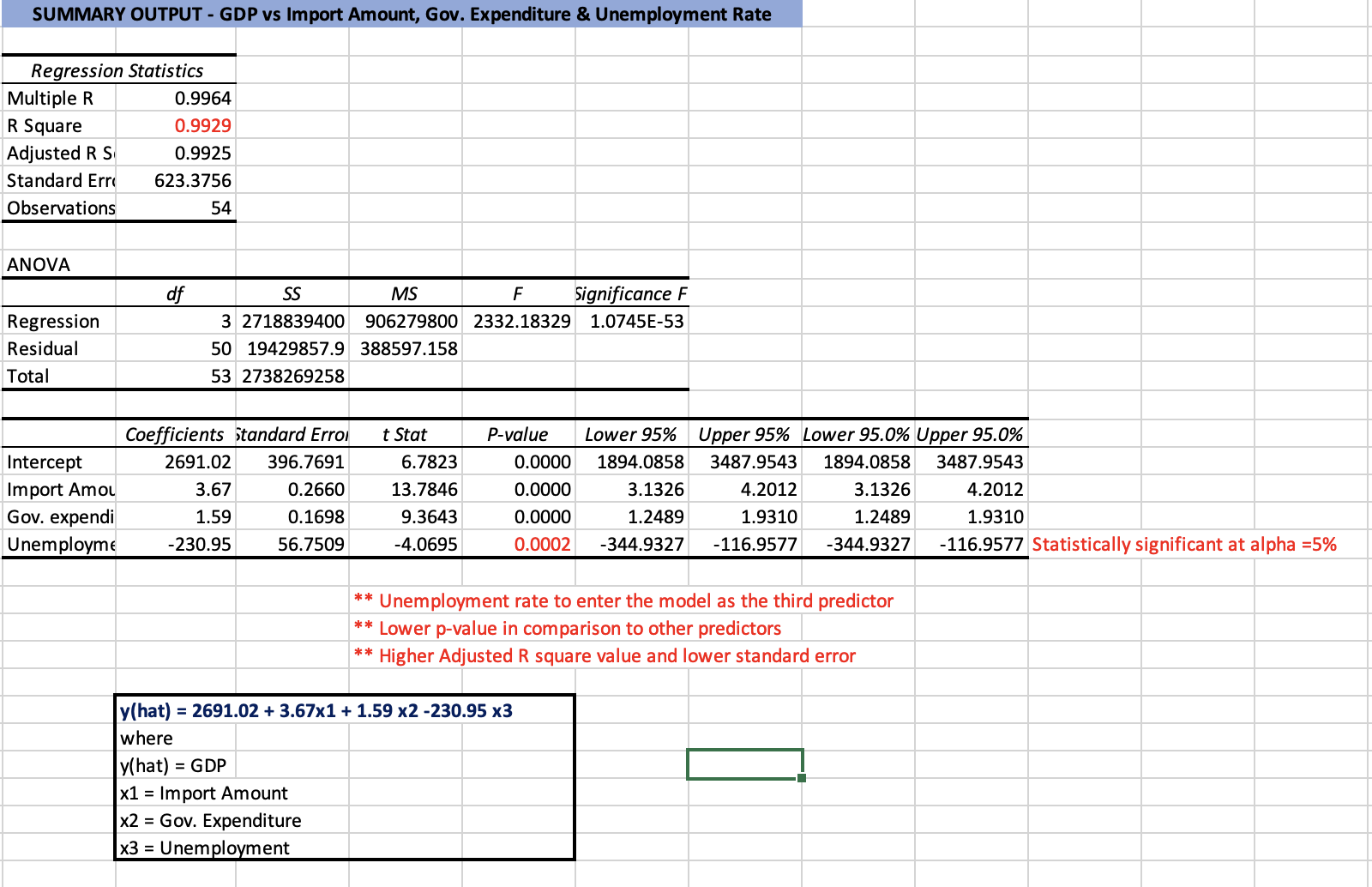
Description automatically generated

* In Step 4, we ran regression for the remainder of the variables, but we could not find any predictor which supported our model significantly. **Hence, Step 3 became our decisive step with the *Import Amount, Government Expenditure and Unemployment Rate* as best predictors for *GDP.***

***A screenshot of a graph

Description automatically generated***

* **Our final regression table:**



**Checking for Assumptions:**

* The following are the assumptions of regression analysis:

1. The model is linear

2. The error terms are constant variances

3. The error terms are independent

4. The error terms are normally distributed

* **Assumption 1, 2 and 3 –** For this**,** we plotted Residuals individually with all the three predictor values to understand if the assumptions are holding true or not.

* From the above three graphs, we are seeing a constant variance of the errors, as do not seem to diverging or converging. **Assumption 2 holds true.**
* There seems to be a loose pattern amongst the error terms. We calculated correlation of the residuals amongst themselves, which came out to be 0.526. It hints as our error terms having a medium strength pattern. **Assumption 3 does not hold true.**
* From these three graphs, we can also say our model appears to be linear. Assumption 1 holds true.
* **Assumption 4 –** For this, we create a histogram of residuals:

* Our errors terms seem almost normally distributed. **Assumption 4 holds true.**