Influences of GDP in the U.S

Ginamarie Mastrorilli

Department of Statistics

University of Connecticut

November 14, 2022

Abstract

This study makes an attempt to determine the influence of specific economic factors namely Population, Business Sector Labor Productivity, 10 Year Treasury Constant Maturity Rate, Disposable Personal Income, Unemployment Rate, and Total Primary Energy Consumption on United States GDP. The data is collected from Federal Reserve Bank of St. Louis and the U.S. Energy Information Administration from the period of January 1st, 1973 to July 1st, 2022. This study's scope is limited to these listed variables. Ordinary Least Squares Regression is used to analyze the relationship between GDP and these variables. The main variable of interest is Total Primary Energy Consumption. The study revealed that Business Sector Labor Productivity, 10 Year Treasury Constant Maturity Rate, Disposable Personal Income, and Unemployment Rateare factors that significantly predict GDP of the United States. There was no significant relationship shown between Total Primary Energy Consumption and GDP in the United States based off of this study.

Introduction

What is Gross Domestic Product? Gross Domestic Product is one of the main factors that goes into determining a countries economic growth. GDP is the total monetary value of all goods and services that are produced in a country and is a "comprehensive measure of U.S economic activity" (Bureau of Economic Analysis (2022)). Gross Domestic Product was originally invented in the 1600's but evolved into governmental use in the 1900's. GDP became a national tool to measure a countries economic activity in the 1940's after the Bretton Woods confrence in New Hampshire, US. At this time, Gross National Product was still a main tool to measure production, but in 1991, the United States swiched to using GDP as its main estimate. GDP is calculated by the equation:

$$C + I + G + NX = GDP \tag{1}$$

In this equation, "(C) represents private-consumption expenditures by households and nonprofit organizations, investment (I) refers to business expenditures by businesses and home purchases by households, government spending (G) denotes expenditures on goods and services by the government, and net exports (NX) represents a nation's exports minus its imports" Britannica (2017). This equation is one that is learned in any introduction to macroeconomics courses. Calculating GDP is something that economists have accomplished and this equation is accepted in the industry. What economists are now researching is what factors impact Gross Domestic Product. There has been an abdunace of research in this field related to what variables are significant. Which factors have been researched are subjective based on those who are conducting the study.

Divya and Devi (2014) found that Exchange rate, Sensex and Balance of Payment reflected by current and capital account balance are important factors that influence India's GDP. Using correlation and ANOVA, they also found that inflation is highly correlated, but not a significant influencer. The main purpose of their study is to identify key signals for an

economy before a crisis occurs.

Van den Bergh (2009) researches the criticism behind GDP's influence in economies. They compare the understanding that even though GDP can influence economically relevant decisions, it does not factor in social welfare. Since GDP is a global indicator 'The importance of GDP information for firms, investors and citizens/consumers is illustrated by the media – television, radio, newspapers, financial and other magazines, and internet – informing us on a daily basis about the status of our national GDP, both over time and in comparison with other countries.'(Van den Bergh (2009)) They also touch on how government agencies and politicians strive to avoid low GDP. When GDP is low, this can lead to negative voter response, and less public expenditures, which are both fatal prospects for those in power(Van den Bergh (2009)).

Szustak et al. (2021) investigated the possible relationship between energy production and GDP growth. Their goal was to discover the direction of the relationship if one exists. found that the relationship between power production and GDP is random. The study found that there is not a relationship between energy production and GDP. Even though in some counties the relationship was stronger, they concluded that the relationship was random.

Kalyoncu et al. (2013) focused on the relationship between energy consumption and economic growth in Georgia, Azerbaijan and Armenia from 1995 to 2009. Using the Engle-Granger cointegration and Granger causality tests, they investigated the direction of causality for the purpose of informing policy makers. Their research found that there is 'unidirectional causality from per capita GDP to per capita energy consumption for Armenia.' (Kalyoncu et al. (2013)) The countries they focused their study on all faced low energy supply which is a factor that must be taken into account.

This paper strives to quantify the relation between influencers of GDP. The main focus is to test whether energy consumption has a significant relationship to Gross Domestic Product. This will be conducted using the variable Total Primary Energy Consumption. Since energy data is readily available on a monthly basis, this indicator could be used by economists to

predict GDP on a monthly scale. The goal is to be able to identify signals of crucial moments to allow policy makers to make more informed decisions.

Data

The economic data used in this research was collected from Federal Reserve Economic Data(FRED). This is an online database that is maintained by the Research Department at the Federal Reserve Bank of St. Louis. The variables collected from the FRED website are Gross Domestic Product (Billions of Dollars, Quarterly, Seasonally Adjusted Annual Rate), Population (Thousands, Quarterly, Not Seasonally Adjusted), Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity, Disposable Personal Income (Billions of Dollars, Quarterly, Seasonally Adjusted Annual), Unemployment Rate (Percent, Quarterly, Seasonally Adjusted), and Business Sector: Labor Productivity (Output per Hour) for All Employed Persons.

The variable Total Primary Energy Consumption (Quadrillion Btu) used in this research was collected from the U.S. Energy Information Administration website. The EIA "collects, analyzes, and disseminates independent and impartial energy information to promote sound policymaking, efficient markets, and public understanding of energy and its interaction with the economy and the environment."(EIA)

In the original dataframe there are 199 observations from 7 variables with no missing values. Observations are complied quarterly from January 1st, 1973 to July 1st, 2022. For this paper, the original dataset will be split into two datasets, a training set and a validation set. This procedure is used to measure the performance of the models built. The training dataset is used to fit the desired model, whereas the validation dataset is used to evaluate the fit of that desired model. For this purpose, the data will be split into .8 train and .2 into validation. The independent variables are Population (POP), Business Sector Labor Productivity (BSLP), 10 Year Treasury Constant Maturity Rate (TCMR),

Disposable Personal Income (DPI), Unemployment Rate (UNRATE), and Total Primary Energy Consumption (TPEC). The dependent variable is Gross Domestic Product (GDP). Summary statistics for the training dataset are as follows:

Summary						
Variable	Min.	1st Q	Median	Mean	3rd Q	Max.
TPEC	5.669	6.07	7.609	7.462	8.082	9.965
GDP	1377	4341	8363	9962	14940	25248
POP	211192	238482	271709	273195	30886	332940
TCMR	.6506	3.4607	6.1448	6.1156	8.054	14.8384
DPI	968.7	3114.2	6037.2	7440.5	11171.3	19586.5
UNRATE	3.6	5.05	5.867	6.231	7.3	12.967
BSPL	45.99	56.35	67.69	75.44	98.96	116.10

Methods

To investigate which factors influence Gross Domestic Product, Ordinary Least Squares Regressiion will be used. Specifically, the variables Population, Business Sector Labor Productivity, 10 Year Treasury Constant Maturity Rate, Disposable Personal Income, Unemployment Rate, and Total Primary Energy Consumption will be used to predict United States Gross Domestic Product.

Ordinary Least Squares Regressiion is a continuation of simple linear regression. For this model, there must be one or more independent predictors that are used to predict one dependent variable. The general form of the OLS regression model is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon \tag{2}$$

In this equation, Y is an outcome that is a continuous measurement. β_0 is the intercept, $\beta_1 X_1$ is the regression coefficent of the first independent variable. $\beta_n X_n$ is the last regression

coefficent for the final variable in the model. ϵ represents how much variation there is in our estimate for the dependent variable, or otherwise known as model error.

OLS Regression Assumptions: - normality - heteroscedacity

AV Plors will also be ised in this analysis to "A strong linear relationship in the added variable plot indicates the increased importance of the contribution of X to the model already containing the other predictors.

Application

When applying OLS regression to the data set, all independent variables were included in the intital model. Model 1 was built using the variables TPEC, POP, TCMR, DPI, UNRATE and BSLP from the training dataset. The equation used for this model is as follows:

$$GDP = \beta_0 + \beta_1 TPEC + \beta_2 TCMR + \beta_3 DPI + \beta_4 UNRATE + \beta_5 BSPL + \beta_6 POP \quad (3)$$

Summary statistics for this model are as follows:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1934.41  -141.24    14.71   116.92  1606.24

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.296e+03  8.253e+02  -6.417 1.67e-09
TPEC        -1.687e+01  4.458e+01  -0.378   0.7057
POP          1.174e-02  5.199e-03   2.257   0.0254
TCMR         8.687e+01  1.932e+01   4.497 1.36e-05
DPI          8.699e-01  3.853e-02  22.575  < 2e-16
UNRATE      -1.937e+02  1.948e+01  -9.944  < 2e-16
BSLP         8.459e+01  1.420e+01   5.957 1.72e-08

(Intercept) ***
TPEC
POP         *
TCMR        ***
DPI         ***
UNRATE      ***
BSLP        ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 344.5 on 152 degrees of freedom
Multiple R-squared:  0.9973,    Adjusted R-squared:  0.9972
F-statistic:  9488 on 6 and 152 DF,  p-value: < 2.2e-16
```

Figure 1: This figure shows the output from R for the summary statistics for Model 1

From the output of Model 1, it concludes that the model is useful. This is because the P-Value of the F-Test is less than $2.2 * 10^{-16}$. The variables TCMR, DPI, UNRATE and BSLP are significant in this model, whereas the variables TPEC and POP are not significantly differnt from zero at the 5% level of significance. The Multiple R-Squared value indicates that this model is a good fit since 99.73% of the variablity in the data is explated by the model. Model 1 found significant relationships between the frequency of GDP and the variables TCMR, DPI, UNRATE and BSLP. This is due to P-Value associated with these variables is less than .001. Specifically it found a 86.87% increase in the frequency of GDP for every 1% increase in TCMR, a 86.99% increase in the frequency of GDP for every 1% increase in DPI, a -19.37% decrease in the frequency of GDP for every 1% increase in UNRATE, and a 84.59% increase in the frequency of GDP for every 1% increase in BSLP. The Added Variable Plots for the model are as follows:

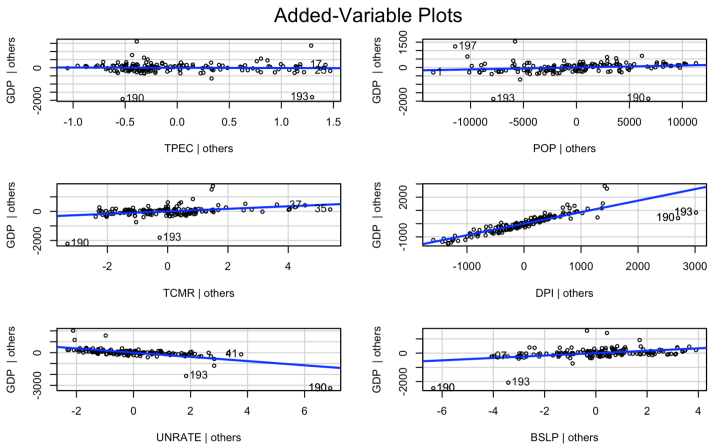

Figure 2: This figure shows the AV Plots for the variables included in Model 1

From the output above, it can be concluded that there is a strong linear relationship between the variables TCMR, DPI, UNRATE and BSLP with GDP. A weak linear relationship is show fromthe varibles TPEC and POP with GDP. This output confirms that the relationship between TPEC and POP with GDP are not significant in this model. To check the assumptions of homoscedacity a residuals vs. fitted plot is shown.

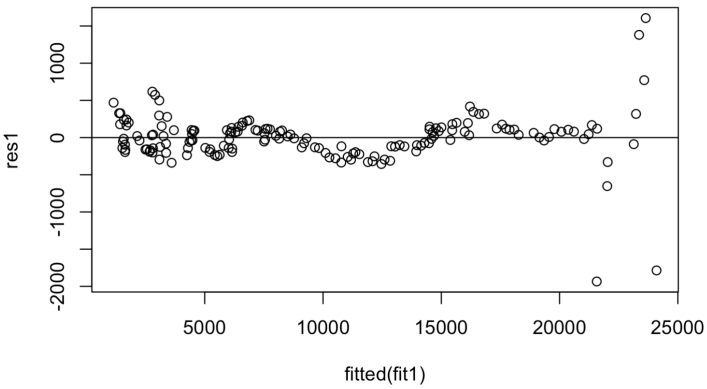

Figure 3: This figure shows the Residuals vs. Fitted plot for Model 1

From this plot, the assumption of homoscedacity is questionable. The residuals do not appear to be randomly scattered throughout the entire plot, and there could be a noticle pattern. This warrents further invesigation. Next, the assumption of normality will be checked using a Q-Q Plot.

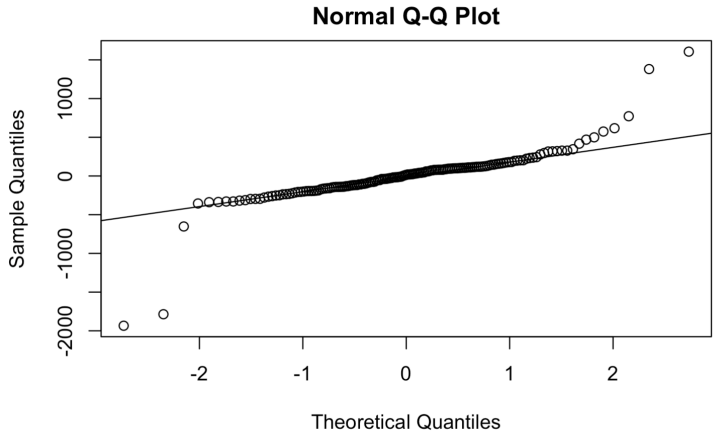

Figure 4: This figure shows the Q-Q Plot for Model 1

From this plot, the assumption of normality is met. The points fall along the line at a rough 45 degree angle. There is a few stray points, but not enough concern that warrents futer invesigation. Therefore, the normality assumption is met.

For the second model, the non significant variables from Model 1 are removed. Model 1 was built using the independent variables TCMR, DPI, UNRATE and BSLP from the training dataset. The equation used for Model 2 is as follows:

$$GDP = \beta_0 + \beta_1 TCMR + \beta_2 DPI + \beta_3 UNRATE + \beta_4 BSPL \tag{4}$$

Summary statistics for Model 2 are as follows:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1867.23  -175.24    32.29   143.46  1532.46

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.867e+03  5.203e+02  -7.432 6.90e-12 ***
TCMR         9.437e+01  1.910e+01   4.940 2.02e-06 ***
DPI          8.584e-01  3.723e-02  23.057  < 2e-16 ***
UNRATE      -2.065e+02  1.754e+01 -11.778  < 2e-16 ***
BSLP         1.081e+02  9.513e+00  11.359  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 348 on 154 degrees of freedom
Multiple R-squared:  0.9972,    Adjusted R-squared:  0.9972
F-statistic: 1.395e+04 on 4 and 154 DF,  p-value: < 2.2e-16
```

Figure 5: This figure shows the output from R for the summary statistics for Model 2

From the output above of Model 2, it concludes that the model is useful. This is because the P-Value of the F-Test is less than $2.2 * 10^{-16}$. All of the variables in this model (TCMR, DPI, UNRATE and BSLP) are significant at the 5% level of significance. The R-Squared value indicates that this model is a good fit since 99.72% of the variablity in the data is explated by the model. Model 2 found significant relationships between the frequency of GDP and the variables TCMR, DPI, UNRATE and BSLP. This is due to P-Value associated with these variables is less than .001. Specifically it found a 94.37% increase in the frequency of GDP for every 1% increase in TCMR, a 85.84% increase in the frequency of GDP for every 1% increase in DPI, a -20.65% decrease in the frequency of GDP for every 1% increase in UNRATE, and a 10.81% increase in the frequency of GDP for every 1% increase in BSLP. The Added-Variable Plots for variables included in Model 2 are as follows:

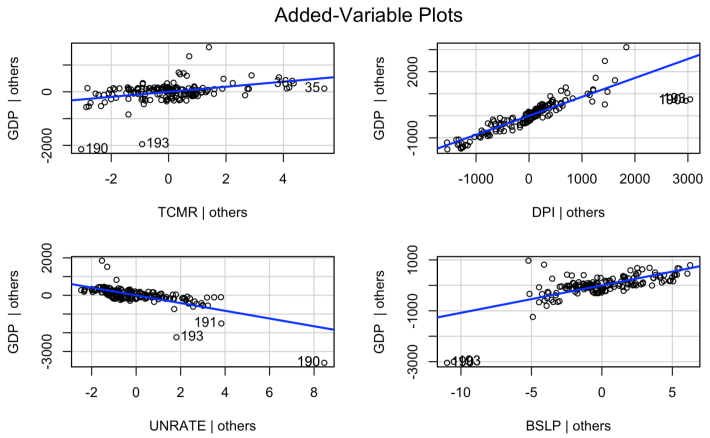

Figure 6: This figure shows the AV Plots for the variables included in Model 2

From the output produced in the Added-Variable Plots from Model 2, it confirms there is a strong linear relationship between the variables TCMR, DPI, UNRATE and BSLP with GDP. To check the assumptions of homoscedacity a residuals vs. fitted plot is shown.

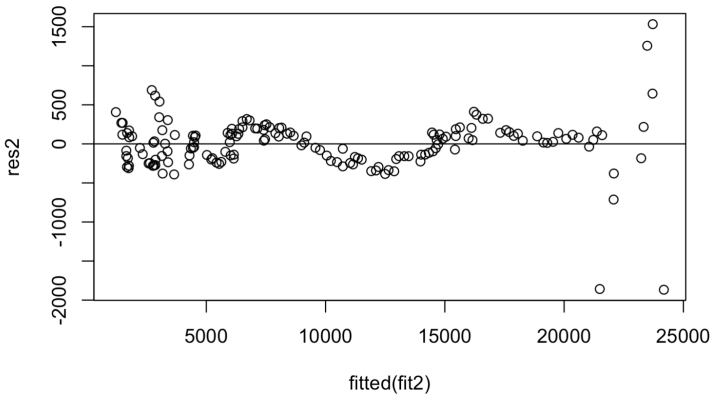

Figure 7: This figure shows the Residuals vs. Fitted plot for Model 2

From this plot, the assumption of homoscedacity is questionable The residuals do not

appear to be randomly scattered throughout the entire plot, and there could be a noticle pattern in some parts of the plot. This warrents further invesigation. Next, the assumption of normality will be checked using a Q-Q Plot.

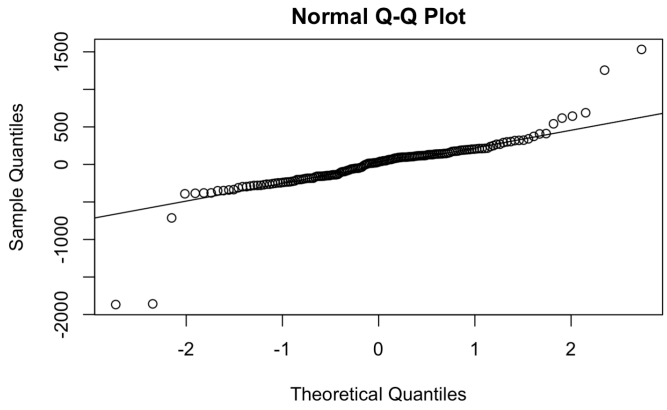

Figure 8: This figure shows the Q-Q Plot for Model 2

From this plot, the assumption of normality is met. The points fall along the line at a rough 45 degree angle. There is a few stray points, but not enough concern that warrents futer invesigation. Therefore, the normality assumption is met.

Next, the validation data will be applied to Model 1 and Model 2. The summary statistics for Model 1 application with Validation data are as follows:

```
Call:
lm(formula = GDP ~ TPEC + POP + TCMR + DPI + UNRATE + BSLP, data = validation)

Residuals:
    Min      1Q  Median      3Q     Max
-690.08  -76.12    3.16   75.55  876.43

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.035e+01  1.344e+03  -0.023   0.9821
TPEC         6.670e+01  6.247e+01   1.068   0.2934
POP         -7.943e-03  9.649e-03  -0.823   0.4163
TCMR         4.572e+01  2.665e+01   1.716   0.0956 .
DPI          1.215e+00  7.250e-02  16.755   <2e-16 ***
UNRATE      -8.408e+01  3.676e+01  -2.287   0.0287 *
BSLP         3.831e+01  2.787e+01   1.375   0.1785
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 237.8 on 33 degrees of freedom
Multiple R-squared:  0.9989,    Adjusted R-squared:  0.9987
F-statistic:  4992 on 6 and 33 DF,  p-value: < 2.2e-16
```

Figure 9: This figure shows the summary statistics for Model 1 application to the validation dataset.

The summary statistics for Model 2 application with Validation data are as follows:

```
Call:
lm(formula = GDP ~ +TCMR + DPI + UNRATE + BSLP, data = validation)

Residuals:
    Min      1Q  Median      3Q     Max
-644.82  -92.70    9.59   72.37  934.10

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -919.82242  764.12225  -1.204  0.23676
TCMR          50.18851   25.33698   1.981  0.05551 .
DPI            1.20921    0.06487  18.640  < 2e-16 ***
UNRATE       -81.67410   29.04954  -2.812  0.00803 **
BSLP          28.06281   16.17421   1.735  0.09153 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 235.5 on 35 degrees of freedom
Multiple R-squared:  0.9989,    Adjusted R-squared:  0.9987
F-statistic:  7634 on 4 and 35 DF,  p-value: < 2.2e-16
```

Figure 10: This figure shows the summary statistics for Model 2 application to the validation dataset.

Next, and AIC test will be applied to Model 1 and Model 2 application to validation

data. This tested resulted in Model 1 reciving and AIC of 559.5405, and Model 2 557.1234. Since Model 2 recived a lower AIC it has the most adequate fit with the least variables when built using validation data. The regression equation results are as follows:

$$GDP = -919.82242 + 50.18851(TCMR) + 1.20921(DPI) - 81.67410(UNRATE) + 28.06281(BSPL) \quad (5)$$

Discussion

- violate homoscedacty assumption

References

Britannica, B. D. (2017). Gross domestic product. *https://www.britannica.com/topic/gross-domestic-product*.

Bureau of Economic Analysis, B. (2022). Gross domestic product. *U.S. Bureau of Economic Analysis (BEA) https://www.bea.gov/data/gdp/gross-domestic-product*.

Divya, K. H. and V. R. Devi (2014). A study on predictors of gdp: Early signals. *Procedia Economics and Finance 11*, 375–382.

Kalyoncu, H., F. Gürsoy, and H. Göcen (2013). Causality relationship between gdp and energy consumption in georgia, azerbaijan and armenia. *International Journal of Energy Economics and Policy 3*(1), 111–117.

Szustak, Grażyna, P., W. Gradoń, and Ł. Szewczyk (2021). The relationship between energy production and gdp: Evidence from selected european economies. *Energies 15*(1), 50.

Van den Bergh, J. C. (2009). The gdp paradox. *Journal of economic psychology 30*(2), 117–135.