Econometrics 322 Lab #7

Self-Identified Problem

Prof. Paczkowski

Enter your Name in the Next Cell

Derek Cheung (RUID 204005689)

Grading Rubric

Score:___: Max(0, 20 - Total Deductions)

Content Area	Deduction	Times Deducted	Check	Comments
Abstract				
Missing	5		[ ]	
Insufficient/Wrong Focus	1		[ ]	
Data Dictionary (Metadata)				
Missing	5		[ ]	
Insufficient/Wrong Form or Wording	1		[ ]	
Graphs				
Missing	5		[ ]	
Missing Title	1		[ ]	
Missing/Wrong Labels	1		[ ]	
Pre-Lab				
Missing	5		[ ]	
Insufficient/Wrong Answer	2 Each		[ ]	
No/Incorrect/Insufficient Model Specification	2		[ ]	
No/Incorrect Statistical Hypothesis Statement	2 Each		[ ]	
Post-Lab				
Missing	5		[ ]	
Insufficient/Wrong Answer	2 Each		[ ]	
Correlations				
Missing	5		[ ]	
Insufficient/Wrong Analysis	2		[ ]	
Missing Graph	2		[ ]	
Estimations				
Missing	5		[ ]	
No or incorrect discussion/interpretation of...				
Hypothesis tests and p-values	2 Each		[ ]	
R^2	2		[ ]	
F-Statistic	2		[ ]	
Multicollinearity/VIF	2		[ ]	
Heteroskedasticity/Test	2		[ ]	
Autocorrelation/Test	2		[ ]	
No/insufficient model selection	2		[ ]	
Elasticities				
Missing	5		[ ]	
Incorrect Interpretation	2		[ ]	
Missing Summary Table	2		[ ]	
Model Portfolio			[ ]	
Missing	5		[ ]	
General Comments:				

Contents

Collaboration Policy

Back to Contents

The submitted assignment must be your work.

> **There is to be no collaboration for this assignment.**

Introduction

Back to Contents

Purpose

Back to Contents

The purpose of this lab is to allow you to:

1. identify and define your own economic problem;
2. collect your own data;
3. analyze your data using the tools you learned;
4. estimate a multiple linear regression model in Pandas;
5. interpret key statistics;
6. identify shortcomings in the proposed linear model;
7. summarize the regression output;
8. estimate elasticities and judge their reasonableness;
9. build a model portfolio;
10. interpret the model results.

Assignment

Back to Contents

You have to define your own problem for this lab. There are two rules:

1. There must be a minimum of five (5) legitimate independent variables. Dummies for a concept variable are not legitimate but the concept variable is legitimate.
2. You cannot repeat a problem discussed in class, used in a lesson, a tutorial, or was in a previous lab.

Problem Statement

Back to Contents

State your problem, why it's a problem, and what you expect to show.

The Chinese Communist Party (CCP) has announced it's economic stimulus plan in response to the Covid-19 pandemic. The CCP has put in place plans to increase government spending in order to build new and improve existing infrastructure. This plan will stimulate the short term economy by creating jobs while being able to benefit in the long term, from a more efficient infrastructure after the pandemic is over. These projects have caused the demand for raw materials (example being construction metals) to increase drastically and has led the prices for scrap metal to record levels. Global shipping companies, being some of the most devastated companies from the Covid-19 pandemic, are looking to cash in on these scrap metal prices by selling some of their older and less fuel efficient industrial cargo ships (ships that would have otherwise continued to sail for much longer periods of times) for a short term injection of cash to hopefully be able to ride out the remainder of the economic slowdown we have experienced since March of 2020. This may cause a problem for the future global economy, as we have become so reliant on the merchant marine fleets; "the unsung heroes of globalization". This decrease in shipping capacity may indicate a slow down in globalization as countries look to become more self sufficient, moving forward after the pandemic. This would then have huge impacts on the global economy as the productivity and efficiency offered by specializtion will definitely be lessened.

In this lab we look to explore this relationship between global gross domestic product and the total deadweight tons of the global merchant fleet (totaly shipping capacity) and try to capture the affect total shipping capacity will have on the global economy. To do this we will estimate a complex regression model that will also take into account world population, global unemployment, global energy consumption, and world literacy rate.

The idea for this lab is from Economics Explained's most recent Youtube video uploaded on 12/03/2020. Here is a link to the video https://www.youtube.com/watch?v=qo-2gDg-37w.

Documentation

Back to Contents

Appropriate documentation.

Abstract

Back to Contents

In this lab we explore the relationship world population, total deadweight tons of the global merchant fleet (total shipping capacity), total global energy consumption, total global unemployment rate, and world literacy rates has on global gross domestic product. Specifically, we try to create an OLS model that tries to predict global gross domestic product based on world population, total deadweight tons of the global merchant fleet (total shipping capacity), total global energy consumption, total global unemployment rate, and world literacy rates. We ultimately conclude that the fit is statistically significant with total global energy consumption having the highest affect on global gross domestic product, and this is further reinforced by the consistent statistical significance of total global energy consumption across all models developed. This leads us to conclude that total global energy consumption is definitely related to global gross domestic product in some degree.

Data Dictionary

Back to Contents

Variable	Values	Source	Mnemonic
Year	1991-2019		Year
Global Gross Domestic Product	trillions of current $US	The World Bank	GGDP
World Population	billions of people	Worldometers	PopulationScaled
Total Deadweight Tons of Global Merchant Fleet	hundred of millions tons	United Nations Conference on Trade and Development	DWTonsScaled
Total Global Energy Consumption	quad Btu	US Energy Information Administration	Energy
Total Global Unemployment Rate	percentage	The World Bank	Unemployment
World Literacy Rate	percentage	Macrotrends	Literacy

Note: The GGDP and Energy values for 2019 are projected values.

Pre-lab

Back to Contents

Data description, testable hypotheses, statistical hypotheses.

Type of Data

This data is secondary, quantitative, macro time series data collected from a variety of sources. It is quantitative and macro because it contains only numerical data, aggregated over the entire world. Additionally, the data is collected starting from 1991 till 2019.

Testable Hypothesis

A potential testable hypothesis is:

We expect the global gross domestic product to increase as total shipping capcity, world population, energy consumption, and world literacy rises, and when global unemployment rates decline. This is because, intuitively, high shipping capacity, energy consumption, and world literacy rates combined with a low unemployment rate should indicate a higher production output. Additionally as the global population rises, we would require more output in order to sustain growth and we have more labor resources to apply towards producitve output.

Tentative Specific Model

A potential specific model takes the form of:

GGDP = beta0 + (beta1 x PopulationScaled) + (beta2 x DWTonsScaled) + (beta3 x Energy) + (beta4 x Unemployment) + (beta5 x Literacy) + noise

where the parameters beta0 represents the intercept, beta1 represents the affect global population has on global gross domestic product, beta2 represents the affect total shipping capacity has on global gross domestic product, beta3 represents the affect total energy consumption has on global gross domestic product, beta4 represents the affect global unemployment has on global gross domestic product, beta5 represents the affect world literacy rates has on global gross domestic product, and noise repesents the random variations or error associated with statistical models.

Statistical Hypothesis

The statisitcal hypothesis corresponding to the testable hypotethis in regards to the parameters of the model takes the form of:

- H0: beta1 = 0 Ha: beta1 > 0
- H0: beta2 = 0 Ha: beta2 > 0
- H0: beta3 = 0 Ha: beta3 > 0
- H0: beta4 = 0 Ha: beta4 < 0
- H0: beta5 = 0 Ha: beta5 > 0

where the alternative of beta1, beta2, beta3, and beta5 states that global population, total shipping capacity, total energy consumption, and world literacy rate has a positive relationship with global gross domestic product, and the alternative of beta4 states that global unemployment has a negtive relationship with global gross domestic product. This follows from our testable hypothesis where we conluded, from intution, that as the world's population increases we must produce more to sustain the population which leads to increases in productivity and more energy consumption and more shipping capacity. Therefore any increases in unemployment should signal productive slowdowns which would then lead to lower global gross domestic product.

The statisitcal hypothesis corresponding to the testable hypotethis in regards to the comparison of the naive and sophisticated model takes the form of:

- H0: beta1 = beta2 = beta3 = beta4 = beta5 = 0
- Ha: at least one parameter (beta1, beta2, beta3, beta4, beta5) ≠ 0

where the null hypothesis says the naive (or restricted) model is better, and the alternative hypothesis says the sophisticated (or unrestricted) model is better.

Tasks

Back to Contents

Data import, data examination, model estimation, elasticity calculations, portfolio construction. Be sure to include all you learned this semester.

Data Import

In [1]:
```
#import libs

import numpy as np
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
from statsmodels.iolib.summary2 import summary_col
from statsmodels.stats.api import anova_lm
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
#import data and set row index to year

df = pd.read_excel('lab7Data.xlsx', index_col = 'Year')
df.head()
```

Out[2]:

Year	GGDP	Population	DWTons	Energy	Unemployment	Literacy	PopulationScaled	DWTonsScaled
2019	87.698	7713468100	1.988305e+06	620.000	5.395	86.30	7.713468	19.883051
2018	86.357	7631091040	1.937777e+06	598.006	5.392	86.30	7.631091	19.377770
2017	81.229	7547858925	1.868174e+06	583.929	5.570	86.13	7.547859	18.681738
2016	76.336	7464022049	1.811526e+06	575.594	5.670	85.89	7.464022	18.115264
2015	75.199	7379797139	1.753092e+06	568.778	5.638	85.60	7.379797	17.530919

In [3]:
```
#drop redundant columns

df = df[['GGDP', 'PopulationScaled', 'DWTonsScaled', 'Energy', 'Unemployment', 'Literacy']]
df.head()
```

Out[3]:

Year	GGDP	PopulationScaled	DWTonsScaled	Energy	Unemployment	Literacy
2019	87.698	7.713468	19.883051	620.000	5.395	86.30
2018	86.357	7.631091	19.377770	598.006	5.392	86.30
2017	81.229	7.547859	18.681738	583.929	5.570	86.13
2016	76.336	7.464022	18.115264	575.594	5.670	85.89
2015	75.199	7.379797	17.530919	568.778	5.638	85.60

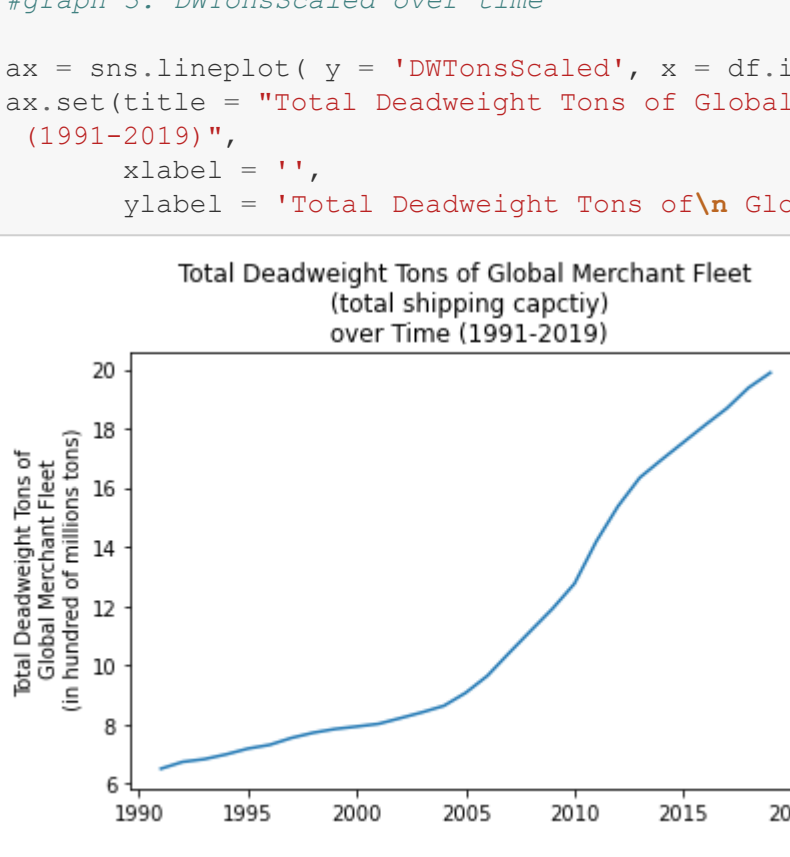
In [4]:
```
#descriptive stats

df.describe().T
```

Out[4]:

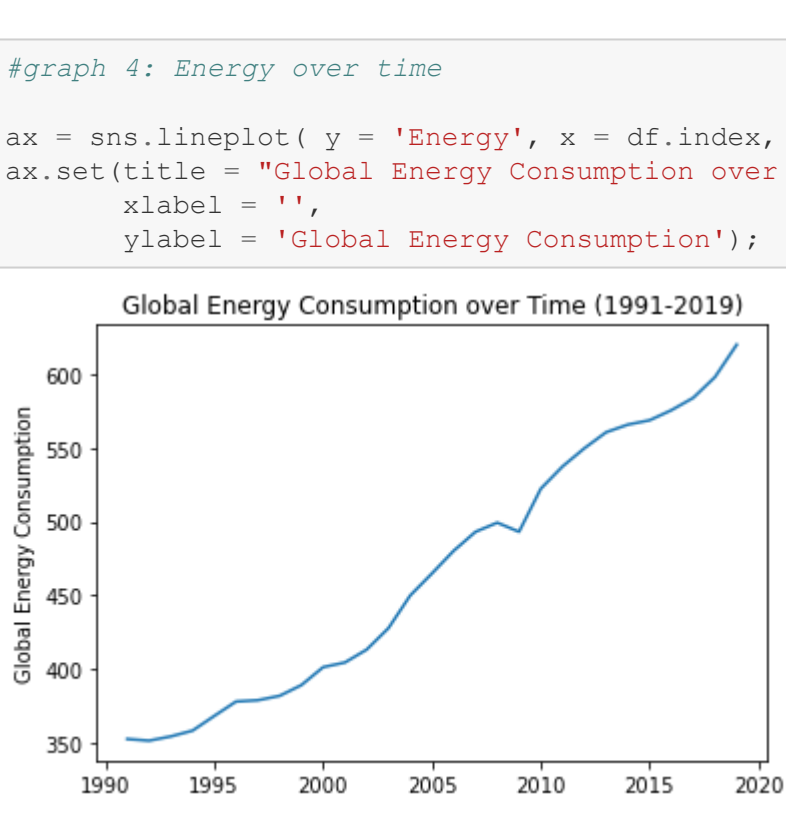
	count	mean	std	min	25%	50%	75%	max
GGDP	29.0	51.892828	21.842192	23.967000	31.573000	47.517000	75.146000	87.698000
PopulationScaled	29.0	6.554847	0.696896	5.414289	5.984794	6.541907	7.125828	7.713468
DWTonsScaled	29.0	11.290307	4.520854	6.512820	7.728012	9.074743	15.374839	19.883051
Energy	29.0	466.414103	87.538966	351.696000	382.076000	465.007000	549.766000	620.000000
Unemployment	29.0	5.645241	0.330158	4.758000	5.430000	5.670000	5.854000	6.189000
Literacy	29.0	81.777586	3.606607	74.910000	80.210000	82.410000	84.730000	86.300000

Data Exploration

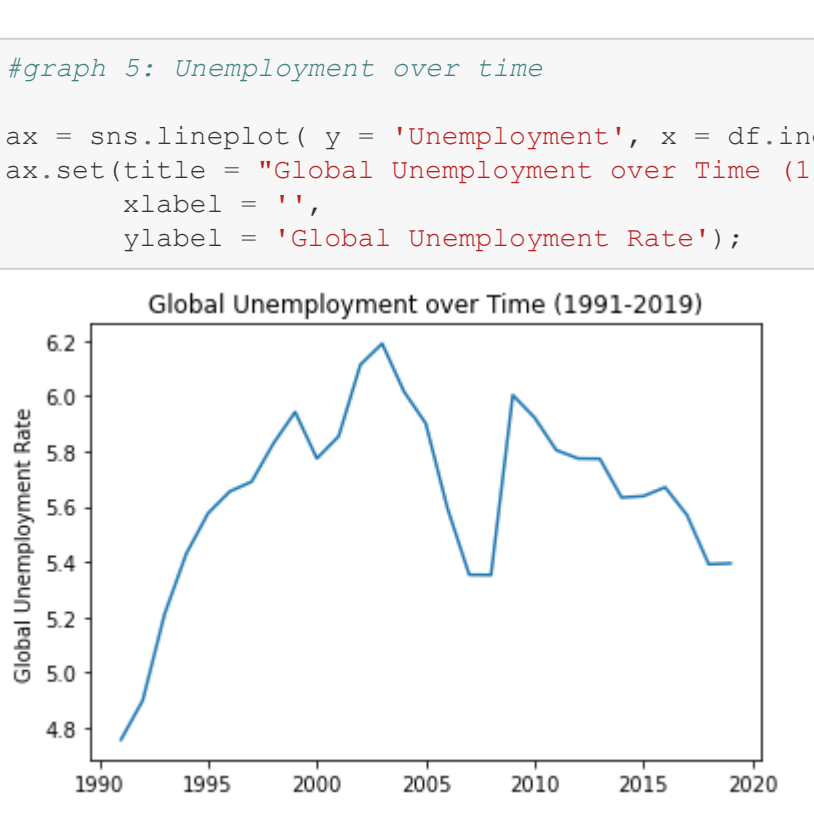
In [5]:
```
#graph 1: GGDP over time

ax = sns.lineplot( y = 'GGDP', x = df.index, data = df);
ax.set(title = "Global Gross Domestic Product over Time (1991-2019)",
       xlabel = '',
       ylabel = 'Global Gross Domestic Product');
```



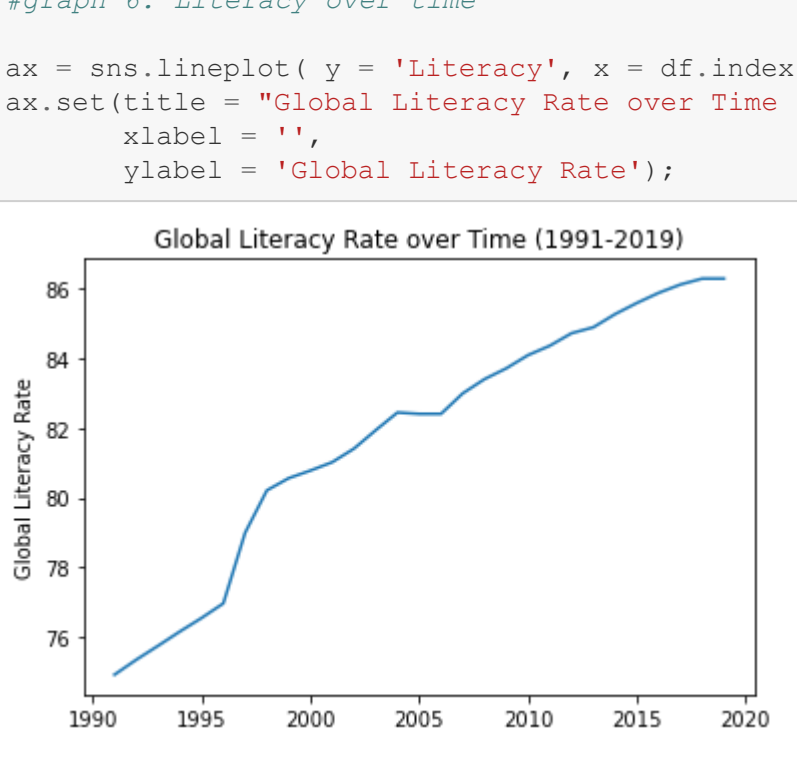
In [6]:
```
#graph 2: Population over time

ax = sns.lineplot( y = 'PopulationScaled', x = df.index, data = df);
ax.set(title = "Global Population over Time (1991-2019)",
       xlabel = '',
       ylabel = 'Global Population (in billions)');
```



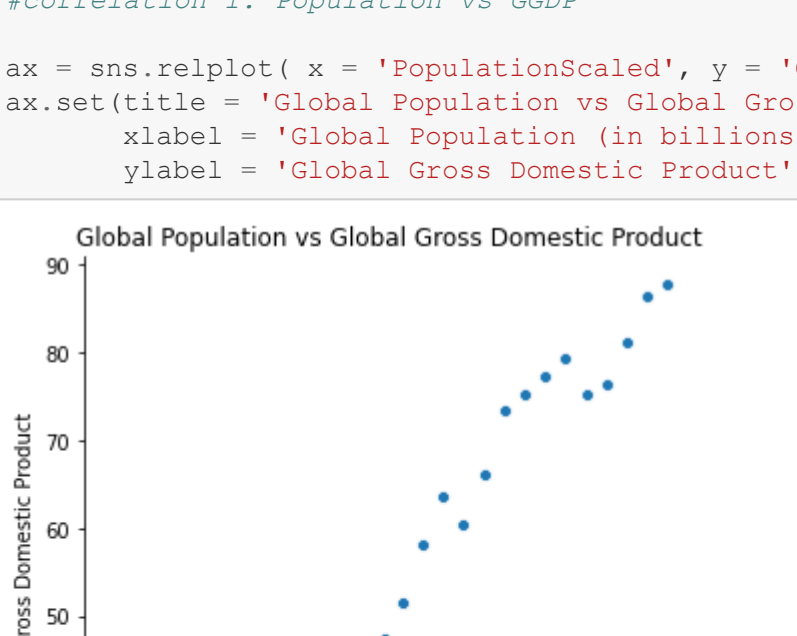
In [7]:
```
#graph 3: DWTonsScaled over time

ax = sns.lineplot( y = 'DWTonsScaled', x = df.index, data = df);
ax.set(title = "Total Deadweight Tons of Global Merchant Fleet\n (total shipping capctiy)\n over Time
 (1991-2019)",
       xlabel = '',
       ylabel = 'Total Deadweight Tons of\n Global Merchant Fleet\n (in hundred of millions tons)');
```



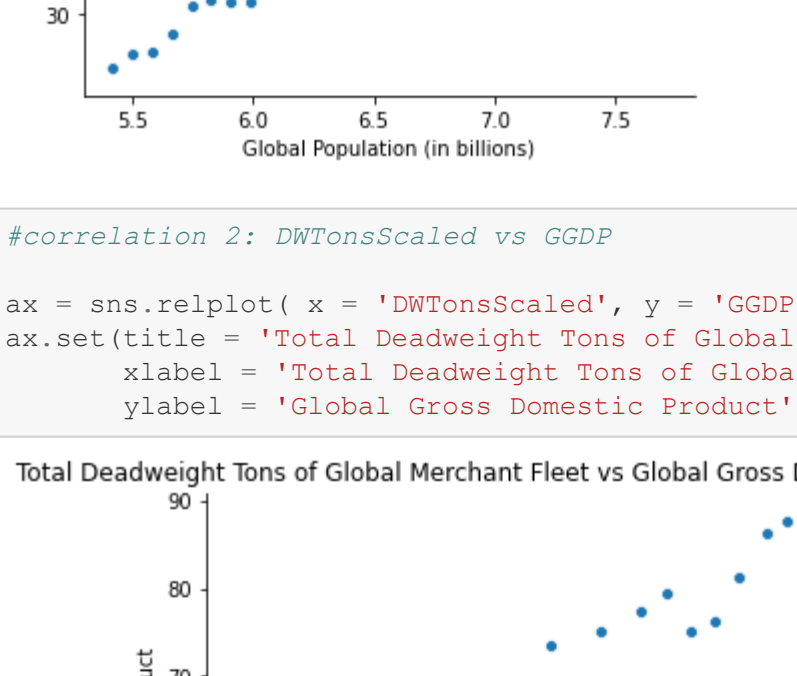
In [8]:
```
#graph 4: Energy over time

ax = sns.lineplot( y = 'Energy', x = df.index, data = df);
ax.set(title = "Global Energy Consumption over Time (1991-2019)",
       xlabel = '',
       ylabel = 'Global Energy Consumption');
```



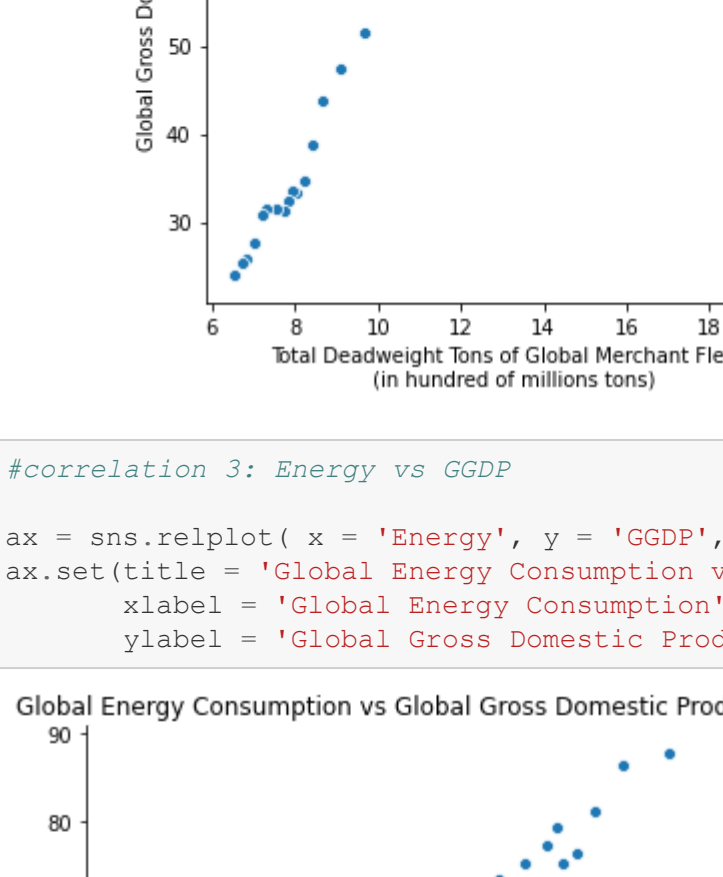
In [9]:
```
#graph 5: Unemployment over time

ax = sns.lineplot( y = 'Unemployment', x = df.index, data = df);
ax.set(title = "Global Unemployment over Time (1991-2019)",
       xlabel = '',
       ylabel = 'Global Unemployment Rate');
```



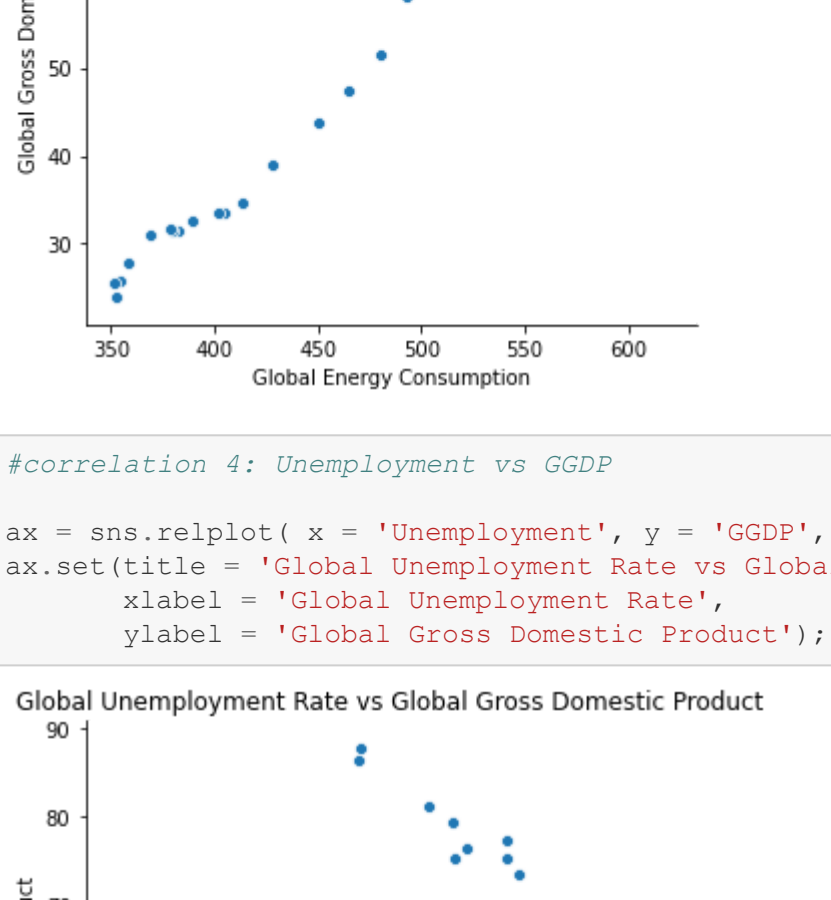
In [10]:
```
#graph 6: Literacy over time

ax = sns.lineplot( y = 'Literacy', x = df.index, data = df);
ax.set(title = "Global Literacy Rate over Time (1991-2019)",
       xlabel = '',
       ylabel = 'Global Literacy Rate');
```



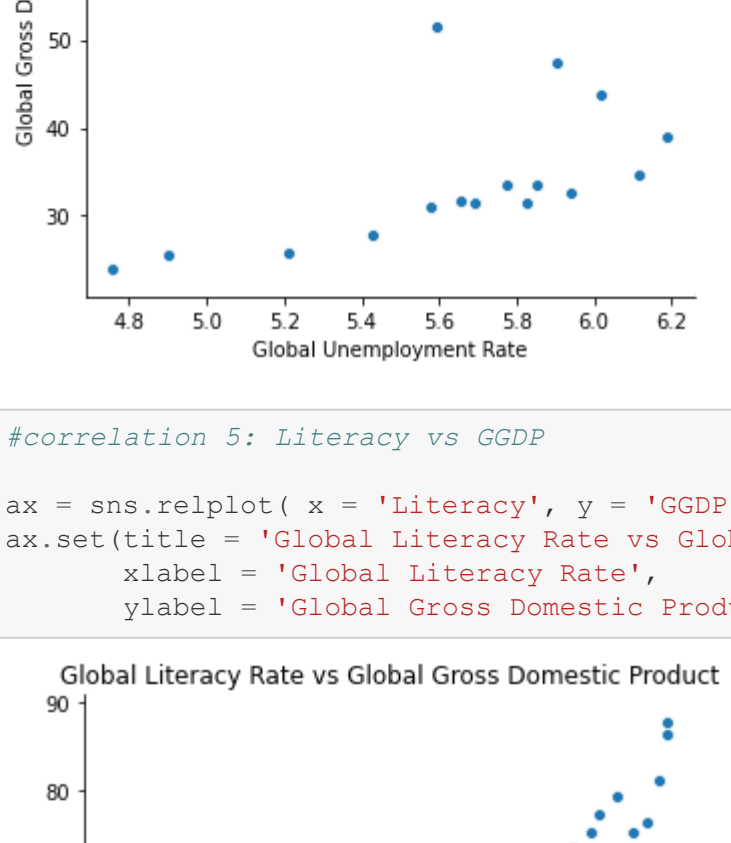
In [11]:
```
#correlation 1: Population vs GGDP

ax = sns.relplot( x = 'PopulationScaled', y = 'GGDP', data = df)
ax.set(title = 'Global Population vs Global Gross Domestic Product',
       xlabel = 'Global Population (in billions)',
       ylabel = 'Global Gross Domestic Product');
```



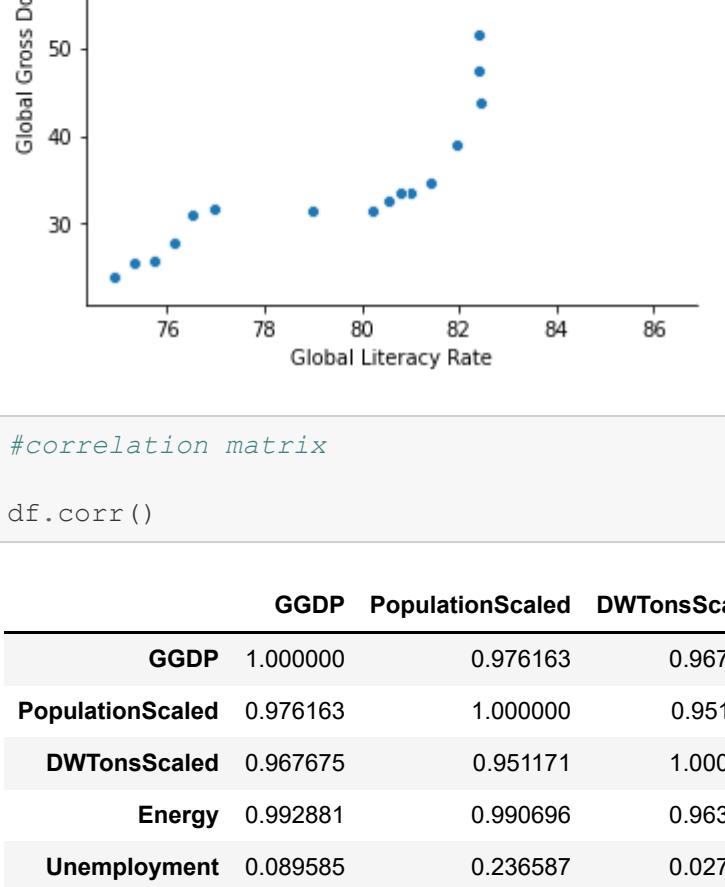
In [12]:
```
#correlation 2: DWTonsScaled vs GGDP

ax = sns.relplot( x = 'DWTonsScaled', y = 'GGDP', data = df)
ax.set(title = 'Total Deadweight Tons of Global Merchant Fleet vs Global Gross Domestic Product',
       xlabel = 'Total Deadweight Tons of Global Merchant Fleet \n (in hundred of millions tons)',
       ylabel = 'Global Gross Domestic Product');
```



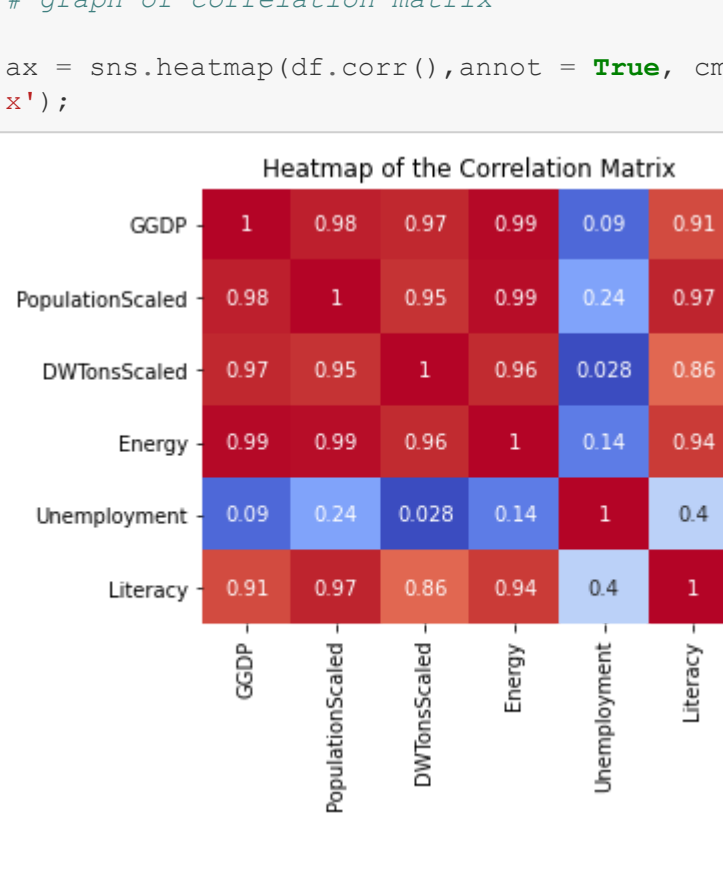
In [13]:
```
#correlation 3: Energy vs GGDP

ax = sns.relplot( x = 'Energy', y = 'GGDP', data = df)
ax.set(title = 'Global Energy Consumption vs Global Gross Domestic Product',
       xlabel = 'Global Energy Consumption',
       ylabel = 'Global Gross Domestic Product');
```



In [14]:
```
#correlation 4: Unemployment vs GGDP

ax = sns.relplot( x = 'Unemployment', y = 'GGDP', data = df)
ax.set(title = 'Global Unemployment Rate vs Global Gross Domestic Product',
       xlabel = 'Global Unemployment Rate',
       ylabel = 'Global Gross Domestic Product');
```



In [15]:
```
#correlation 5: Literacy vs GGDP

ax = sns.relplot( x = 'Literacy', y = 'GGDP', data = df)
ax.set(title = 'Global Literacy Rate vs Global Gross Domestic Product',
       xlabel = 'Global Literacy Rate',
       ylabel = 'Global Gross Domestic Product');
```

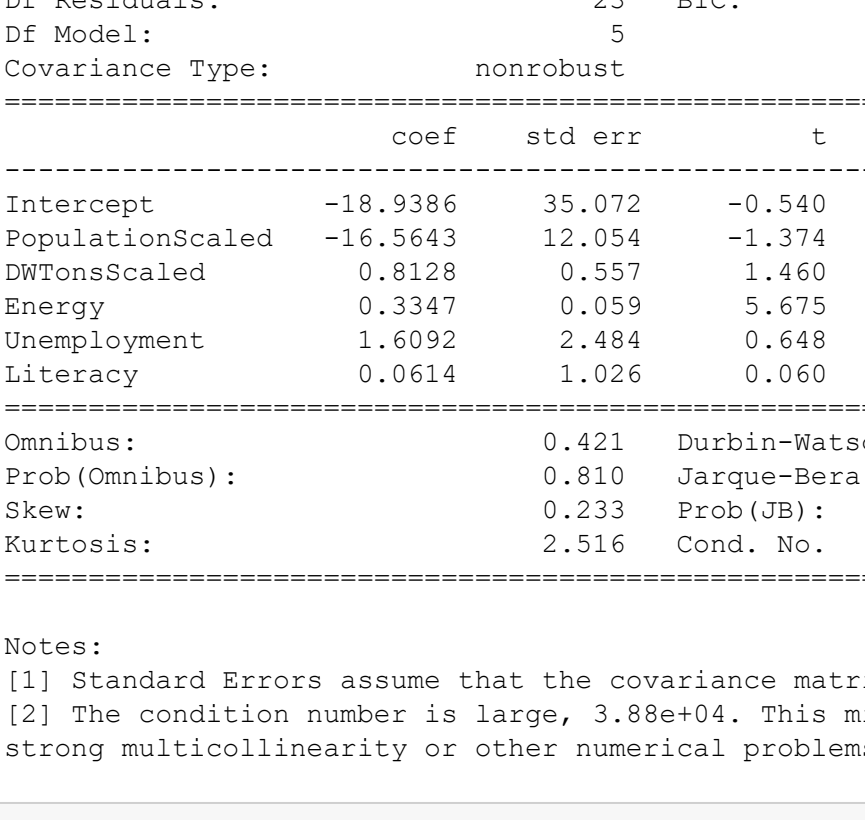


In [16]:
```
#correlation matrix

df.corr()
```

Out[16]:

	GGDP	PopulationScaled	DWTonsScaled	Energy	Unemployment	Literacy
GGDP	1.000000	0.976163	0.967675	0.992881	0.089585	0.913640
PopulationScaled	0.976163	1.000000	0.951171	0.990696	0.236587	0.970719
DWTonsScaled	0.967675	0.951171	1.000000	0.963606	0.027839	0.863517
Energy	0.992881	0.990696	0.963606	1.000000	0.135322	0.940330
Unemployment	0.089585	0.236587	0.027839	0.135322	1.000000	0.401828
Literacy	0.913640	0.970719	0.863517	0.940330	0.401828	1.000000

In [17]:
```
# graph of correlation matrix

ax = sns.heatmap(df.corr(),annot = True, cmap = 'coolwarm').set_title('Heatmap of the Correlation Matri
x');
```



Model(s) Estimation

In [18]:
```
#regression 1: all variables used

mod1 = smf.ols(formula = "GGDP ~ PopulationScaled + DWTonsScaled + Energy + Unemployment + Literacy", d
ata = df)
reg1 = mod1.fit()
print(reg1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   GGDP   R-squared:                       0.990
Model:                            OLS   Adj. R-squared:                  0.988
Method:                 Least Squares   F-statistic:                     470.3
Date:                Sun, 06 Dec 2020   Prob (F-statistic):           2.34e-22
Time:                        16:06:24   Log-Likelihood:                -62.836
No. Observations:                  29   AIC:                             137.7
Df Residuals:                      23   BIC:                             145.9
Df Model:                           5
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept          -18.9386     35.072     -0.540      0.594     -91.490      53.612
PopulationScaled   -16.5643     12.054     -1.374      0.183     -41.500       8.371
DWTonsScaled         0.8128      0.557      1.460      0.158      -0.339       1.964
Energy               0.3347      0.059      5.675      0.000       0.213       0.457
Unemployment         1.6092      2.484      0.648      0.524      -3.530       6.749
Literacy             0.0614      1.026      0.060      0.953      -2.062       2.185
==============================================================================
Omnibus:                        0.421   Durbin-Watson:                   1.016
Prob(Omnibus):                  0.810   Jarque-Bera (JB):                0.545
Skew:                           0.233   Prob(JB):                        0.762
Kurtosis:                       2.516   Cond. No.                     3.88e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.88e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [19]:
```
#regression 2: DWTonsScaled only

mod2 = smf.ols(formula = "GGDP ~ DWTonsScaled", data = df)
reg2 = mod2.fit()
print(reg2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   GGDP   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.934
Method:                 Least Squares   F-statistic:                     397.5
Date:                Sun, 06 Dec 2020   Prob (F-statistic):           1.10e-17
Time:                        16:06:24   Log-Likelihood:                -90.123
No. Observations:                  29   AIC:                             184.2
Df Residuals:                      27   BIC:                             187.0
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept       -0.8923      2.845     -0.314      0.756      -6.730       4.945
DWTonsScaled     4.6753      0.234     19.937      0.000       4.194       5.156
==============================================================================
Omnibus:                        3.622   Durbin-Watson:                   0.217
Prob(Omnibus):                  0.163   Jarque-Bera (JB):                3.013
Skew:                           0.688   Prob(JB):                        0.222
Kurtosis:                       2.225   Cond. No.                         33.3
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [20]:
```
#regression 3: DWTonsScaled, Energy, and Unemployment only

mod3 = smf.ols(formula = "GGDP ~ DWTonsScaled + Energy + Unemployment", data = df)
reg3 = mod3.fit()
print(reg3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   GGDP   R-squared:                       0.988
Model:                            OLS   Adj. R-squared:                  0.987
Method:                 Least Squares   F-statistic:                     716.2
Date:                Sun, 06 Dec 2020   Prob (F-statistic):           2.35e-24
Time:                        16:06:24   Log-Likelihood:                -65.326
No. Observations:                  29   AIC:                             138.7
Df Residuals:                      25   BIC:                             144.1
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept      -45.2592      8.337     -5.428      0.000     -62.431     -28.088
DWTonsScaled     0.4972      0.420      1.183      0.248      -0.369       1.363
Energy           0.2242      0.022     10.233      0.000       0.179       0.269
Unemployment    -2.3060      1.553     -1.485      0.150      -5.505       0.893
==============================================================================
Omnibus:                        4.571   Durbin-Watson:                   0.676
Prob(Omnibus):                  0.102   Jarque-Bera (JB):                1.744
Skew:                           0.153   Prob(JB):                        0.418
Kurtosis:                       1.838   Cond. No.                     8.68e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.68e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [21]:
```
#regression 4: Energy only

mod4 = smf.ols(formula = "GGDP ~ Energy", data = df)
reg4 = mod4.fit()
print(reg3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   GGDP   R-squared:                       0.988
Model:                            OLS   Adj. R-squared:                  0.987
Method:                 Least Squares   F-statistic:                     716.2
Date:                Sun, 06 Dec 2020   Prob (F-statistic):           2.35e-24
Time:                        16:06:24   Log-Likelihood:                -65.326
No. Observations:                  29   AIC:                             138.7
Df Residuals:                      25   BIC:                             144.1
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept      -45.2592      8.337     -5.428      0.000     -62.431     -28.088
DWTonsScaled     0.4972      0.420      1.183      0.248      -0.369       1.363
Energy           0.2242      0.022     10.233      0.000       0.179       0.269
Unemployment    -2.3060      1.553     -1.485      0.150      -5.505       0.893
==============================================================================
Omnibus:                        4.571   Durbin-Watson:                   0.676
Prob(Omnibus):                  0.102   Jarque-Bera (JB):                1.744
Skew:                           0.153   Prob(JB):                        0.418
Kurtosis:                       1.838   Cond. No.                     8.68e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.68e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [22]:
```
#regression 5: restricted model

restrictedModel = smf.ols( formula = 'GGDP ~ 1', data = df )
restrictedReg = restrictedModel.fit()
print( restrictedReg.summary() )
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   GGDP   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                       nan
Date:                Sun, 06 Dec 2020   Prob (F-statistic):                nan
Time:                        16:06:24   Log-Likelihood:                -130.07
No. Observations:                  29   AIC:                             262.1
Df Residuals:                      28   BIC:                             263.5
Df Model:                           0
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     51.8928      4.056     12.794      0.000      43.585      60.201
==============================================================================
Omnibus:                       16.701   Durbin-Watson:                   0.026
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                3.079
Skew:                           0.246   Prob(JB):                        0.214
Kurtosis:                       1.481   Cond. No.                         1.00
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Multicolinearity and Autocorrelation Corrections

The best model so far is reg1 based on AIC score. The corrections will be performed on reg1.

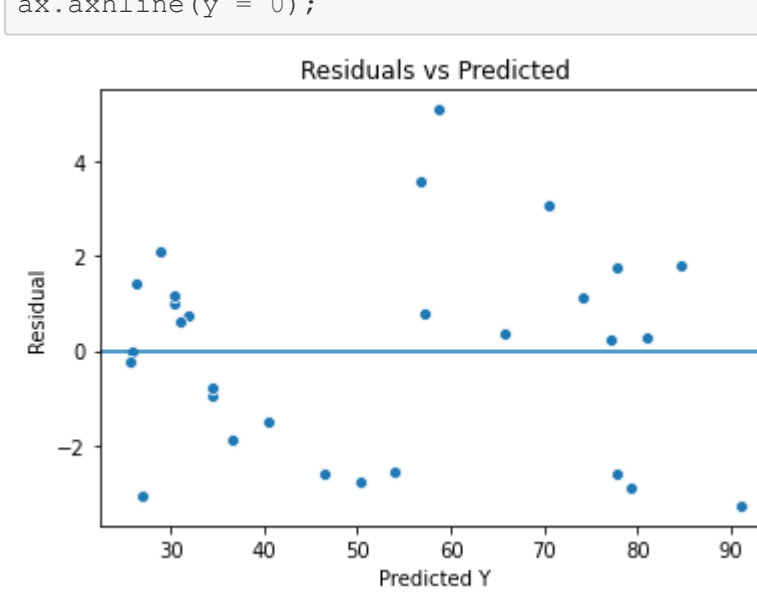
In [23]:
```
#retrieve X variables

X = reg1.model.data.orig_exog
```

In [24]:
```
#add residuals and predicted values

X['Residuals'] = reg1.resid
X['Predict'] = reg1.predict()
X.head()
```

Out[24]:

Year	Intercept	PopulationScaled	DWTonsScaled	Energy	Unemployment	Literacy	Residuals	Predict
2019	1.0	7.713468	19.883051	620.000	5.395	86.30	-3.272028	90.970028
2018	1.0	7.631091	19.377770	598.006	5.392	86.30	1.800000	84.557000
2017	1.0	7.547859	18.681738	583.929	5.570	86.13	0.295035	80.933965
2016	1.0	7.464022	18.115264	575.594	5.670	85.89	-2.882447	79.218447
2015	1.0	7.379797	17.530919	568.778	5.638	85.60	-2.588798	77.787798

In [25]:
```
#plot residuals vs predicted

ax = sns.scatterplot(x = 'Predict', y = 'Residuals', data = X)
ax.set(title = 'Residuals vs Predicted',
      xlabel = 'Predicted Y',
      ylabel = 'Residual')
ax.axhline(y = 0);
```



In [26]:
```
#white test

X = reg1.model.data.orig_exog
white = sm.stats.diagnostic.het_white(reg1.resid, X)
ret = ['Lagrange Multiplier Test Statistic', 'P-Value', 'F-Statistic', 'P-Value']
xzip01 = zip(ret, white)
print("\nWhite's Test for Heteroskedasticity")
dfWhite = pd.DataFrame(list(xzip01), columns = ['stats', 'values'])
dfWhite
```

```
White's Test for Heteroskedasticity
```

Out[26]:

	stats	values
0	Lagrange Multiplier Test Statistic	2.900000e+01
1	P-Value	3.608992e-01
2	F-Statistic	1.542985e+20
3	P-Value	6.364128e-11

In [27]:
```
#regression 6: all variable w correction for heteroskedasticty

mod6 = smf.ols(formula = "GGDP ~ PopulationScaled + DWTonsScaled + Energy + Unemployment + Literacy", d
ata = df)
reg6 = mod6.fit(cov_type = 'HC1')
print(reg6.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   GGDP   R-squared:                       0.990
Model:                            OLS   Adj. R-squared:                  0.988
Method:                 Least Squares   F-statistic:                     616.4
Date:                Sun, 06 Dec 2020   Prob (F-statistic):           1.07e-23
Time:                        16:06:24   Log-Likelihood:                -62.836
No. Observations:                  29   AIC:                             137.7
Df Residuals:                      23   BIC:                             145.9
Df Model:                           5
Covariance Type:                  HC1
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept          -18.9386     26.224     -0.722      0.470     -70.336      32.459
PopulationScaled   -16.5643     11.476     -1.443      0.149     -39.057       5.929
DWTonsScaled         0.8128      0.617      1.316      0.188      -0.397       2.023
Energy               0.3347      0.052      6.419      0.000       0.233       0.437
Unemployment         1.6092      2.945      0.546      0.585      -4.163       7.381
Literacy             0.0614      0.852      0.072      0.943      -1.608       1.731
==============================================================================
Omnibus:                        0.421   Durbin-Watson:                   1.016
Prob(Omnibus):                  0.810   Jarque-Bera (JB):                0.545
Skew:                           0.233   Prob(JB):                        0.762
Kurtosis:                       2.516   Cond. No.                     3.88e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 3.88e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Interpretation of Heteroskedasticity Correction

The correction has yielded an identical model to reg1 where there was no heteroskedasticity correction. This may indicate that the data is homoskedastic.

In [28]:
```
#regression 7: all variables w correction for autocorrelation

mod7 = sm.GLSAR.from_formula(formula = "GGDP ~ PopulationScaled + DWTonsScaled + Energy + Unemployment
 + Literacy",
                              rho = 1,
                              data = df)
reg7 = mod7.iterative_fit(maxiter = 3)
print(reg7.summary())
```

```
                           GLSAR Regression Results
==============================================================================
Dep. Variable:                   GGDP   R-squared:                       0.962
Model:                          GLSAR   Adj. R-squared:                  0.954
Method:                 Least Squares   F-statistic:                     112.7
Date:                Sun, 06 Dec 2020   Prob (F-statistic):           6.43e-15
Time:                        16:06:24   Log-Likelihood:                -54.101
No. Observations:                  28   AIC:                             120.2
Df Residuals:                      22   BIC:                             128.2
Df Model:                           5
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept          -39.3084     40.714     -0.965      0.345    -123.745      45.128
PopulationScaled     1.2255     13.466      0.091      0.928     -26.701      29.151
DWTonsScaled         0.9619      0.755      1.274      0.216      -0.603       2.527
Energy               0.2061      0.067      3.093      0.005       0.068       0.344
Unemployment        -2.8268      2.638     -1.072      0.295      -8.297       2.643
Literacy            -0.0908      1.083     -0.084      0.934      -2.338       2.156
==============================================================================
Omnibus:                        0.062   Durbin-Watson:                   1.474
Prob(Omnibus):                  0.969   Jarque-Bera (JB):                0.281
Skew:                           0.006   Prob(JB):                        0.869
Kurtosis:                       2.509   Cond. No.                     2.22e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.22e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Interpretation of Autocorrellation Correction

The Durbin-Watson statistic prior to the autocorrelation correction was 1.016, indicating there was a degree of autocorrelation. After performing the correction, the Durbin-Watson statstic improve to 1.474, indicating there is still a degree of autocorrelation in the model.

Elasticity Calculations

In [29]:
```
#elasticities for population

slopes = [reg1.params[1], reg6.params[1], reg7.params[1]]
mean_ray = df.GGDP.mean()/df.PopulationScaled.mean()
eta = np.array( slopes ) * mean_ray
print('Estimated Slopes: ', slopes)
print('Elasticities: ', eta)
print('Mean of Population =', round(df.PopulationScaled.mean(),4))
```

```
Estimated Slopes:  [-16.5643408057883, -16.5643408057883, 1.2254613898816729]
Elasticities:  [-131.13509512 -131.13509512    9.70162337]
Mean of Population = 6.5548
```

In [30]:
```
#elasticities for total shipping capacity

slopes = [reg1.params[1], reg2.params[1], reg3.params[1], reg6.params[1], reg7.params[1]]
mean_ray = df.GGDP.mean()/df.DWTonsScaled.mean()
eta = np.array( slopes ) * mean_ray
print('Estimated Slopes: ', slopes)
print('Elasticities: ', eta)
print('Mean of DWTonsScaled =', round(df.DWTonsScaled.mean(),4))
```

```
Estimated Slopes:  [-16.5643408057883, 4.675256322177113, 0.49722298911172236, -16.5643408057883, 1.2
254613898816729]
Elasticities:  [-76.13349138  21.48854525   2.28535036 -76.13349138   5.63250028]
Mean of DWTonsScaled = 11.2903
```

In [31]:
```
#elasticities for energy

slopes = [reg1.params[1], reg3.params[1], reg4.params[1], reg6.params[1], reg7.params[1]]
mean_ray = df.GGDP.mean()/df.Energy.mean()
eta = np.array( slopes ) * mean_ray
print('Estimated Slopes: ', slopes)
print('Elasticities: ', eta)
print('Mean of Energy =', round(df.Energy.mean(),4))
```

```
Estimated Slopes:  [-16.5643408057883, 0.49722298911172236, 0.24773761173469977, -16.5643408057883,
1.2254613898816729]
Elasticities:  [-1.84293415  0.0553206   0.02756307 -1.84293415  0.13634377]
Mean of Energy = 466.4141
```

In [32]:
```
#elasticities for unemployment

slopes = [reg1.params[1], reg3.params[1], reg6.params[1], reg7.params[1]]
mean_ray = df.GGDP.mean()/df.Unemployment.mean()
eta = np.array( slopes ) * mean_ray
print('Estimated Slopes: ', slopes)
print('Elasticities: ', eta)
print('Mean of Unemployment =', round(df.Unemployment.mean(),4))
```

```
Estimated Slopes:  [-16.5643408057883, 0.49722298911172236, -16.5643408057883, 1.2254613898816729]
Elasticities:  [-152.2646108     4.57062951 -152.2646108    11.26482507]
Mean of Unemployment = 5.6452
```

In [33]:
```
#elasticities for literacy

slopes = [reg1.params[1], reg6.params[1], reg7.params[1]]
mean_ray = df.GGDP.mean()/df.Literacy.mean()
eta = np.array( slopes ) * mean_ray
print('Estimated Slopes: ', slopes)
print('Elasticities: ', eta)
print('Mean of Literacy =', round(df.Literacy.mean(),4))
```

```
Estimated Slopes:  [-16.5643408057883, -16.5643408057883, 1.2254613898816729]
Elasticities:  [-10.51107671 -10.51107671   0.77762942]
Mean of Literacy = 81.7776
```

Elasticity Table

Variable	Model	Estimate	Mean	Elasticity	Interpretation
PopulationScaled	reg1	-16.5643	6.5548	-131.1351	Elastic
DWTonsScaled	reg1	-16.5643	1129.0307	-76.1334	Elastic
Energy	reg1	-16.5643	466.4141	-1.8429	Elastic
Unemployment	reg1	-16.5643	5.6452	-152.2646	Elastic
Literacy	reg1	-16.5643	81.7776	-10.5111	Elastic
DWTonsScaled	reg2	0.0468	1129.0307	21.4885	Elastic
DWTonsScaled	reg3	0.0050	1129.0307	2.2854	Elastic
Energy	reg3	0.0050	466.4141	0.0553	Inelastic
Unemployment	reg3	0.0050	5.6452	4.5707	Elastic
Energy	reg4	0.2477	466.4141	0.0276	Inelastic
PopulationScaled	reg6	-16.5643	6.5548	-131.1351	Elastic
DWTonsScaled	reg6	-16.5643	1129.0307	-76.1334	Elastic
Energy	reg6	-16.5643	466.4141	-1.8429	Elastic
Unemployment	reg6	-16.5643	5.6452	-152.2646	Elastic
Literacy	reg6	-16.5643	81.7776	-10.5111	Elastic
PopulationScaled	reg7	1.22546	6.5548	9.7016	Elastic
DWTonsScaled	reg7	1.22546	1129.0307	5.6325	Elastic
Energy	reg7	1.22546	466.4141	0.1363	Inelastic
Unemployment	reg7	1.22546	5.6452	11.2648	Elastic
Literacy	reg7	1.22546	81.7776	0.7776	Inelastic

Model Portfolio

In [34]:
```
model_names = ['reg1', 'reg2', 'reg3', 'reg4', 'restrictedReg', 'reg6', 'reg7']

info_dict = { '\nn': lambda x: "{0:d}".format( int( x.nobs ) ),
              'R2 Adjusted': lambda x: "{:0.3f}".format( x.rsquared_adj ),
              'AIC': lambda x: "{:0.2f}".format( x.aic ),
              'F': lambda x: "{:0.2f}".format( x.fvalue ),}

summary_table = summary_col( [reg1, reg2, reg3, reg4, restrictedReg, reg6, reg7],
            float_format = '%0.4f',
            model_names = model_names,
            stars = True,
            info_dict = info_dict
)
summary_table.add_title( 'Summary Table for GGDP Models' )
print( summary_table )
```

```
                                  Summary Table for GGDP Models
=================================================================================================
                 reg1      reg2       reg3        reg4     restrictedReg    reg6       reg7
-------------------------------------------------------------------------------------------------
DWTonsScaled     0.8128    4.6753*** 0.4972                                0.8128    0.9619
                 (0.5566)  (0.2345)  (0.4205)                              (0.6175)  (0.7548)
Energy           0.3347***           0.2242***   0.2477***                 0.3347*** 0.2061***
                 (0.0590)            (0.0219)    (0.0057)                  (0.0521)  (0.0666)
Intercept        -18.9386  -0.8923   -45.2592*** -63.6555*** 51.8928***    -18.9386  -39.3084
                 (35.0715) (2.8451)  (8.3375)    (2.7127)    (4.0560)      (26.2239) (40.7143)
Literacy         0.0614                                                    0.0614    -0.0908
                 (1.0263)                                                  (0.8520)  (1.0835)
PopulationScaled -16.5643                                                  -16.5643  1.2255
                 (12.0540)                                                 (11.4762) (13.4656)
R-squared        0.9903    0.9364    0.9885      0.9858      0.0000        0.9903    0.9624
R-squared Adj.   0.9882    0.9340    0.9871      0.9853      0.0000        0.9882    0.9539
Unemployment     1.6092              -2.3060                               1.6092    -2.8268
                 (2.4844)            (1.5534)                              (2.9449)  (2.6377)

n               29        29        29          29          29             29        28
R2 Adjusted      0.988     0.934     0.987       0.985       0.000         0.988     0.954
AIC              137.67    184.25    138.65      140.74      262.14        137.67    120.20
F                470.25    397.50    716.17      1876.07     nan           616.41    112.72
=================================================================================================
Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01
```

Post-lab Analysis and Conclusions

Back to Contents

Apply the analysis tools you learned.

Relationship between GGDP and independent variables analysis and testable hypothesis conclusion

Different models show slightly differing relationships between gross global domestic product and the independent variables. Across all models it seems there is a slight positive relationship between total shipping capactiy and gross global domestic product. Additionally, in terms of statistical significance, the various models consistently show total energy consumption as the most important variable in predicting global gross domestic product.

Correlation matrix interpretation

The correlation matrix shows intense positive correlation between many of the independent variables. For example, gross global domestic product, total shipping capacity, total energy consumption, world population, and global literacy has steadily risen continuously since the beginning of our observation period. This may be because of some other underlying factors I did not take into account that may have introduced a huge source of error due to colinearity.

R^2 and R^2 adjusted interpretation

Reg1 shows a R2 of 0.9903 and a R2-adjusted of 0.9882. This tells us that 99.03% of the variation is described by the reg1 model, and 98.82% of the variation is described by the indepenedent variables that affect global gross domestic product. Reg2 shows a R2 of 0.9364 and a R2-adjusted of 0.9340. This tells us that 93.64% of the variation is described by the reg2 model, and 93.40% of the variation is described by the indepenedent variables that affect global gross domestic product. Reg3 shows a R2 of 0.9885 and a R2-adjusted of 0.9871. This tells us that 98.85% of the variation is described by the reg3 model, and 98.71% of the variation is described by the indepenedent variables that affect global gross domestic product. Reg4 shows a R2 of 0.9858 and a R2-adjusted of 0.9853. This tells us that 98.58% of the variation is described by the reg4 model, and 98.53% of the variation is described by the indepenedent variables that affect global gross domestic product. Reg6 shows a R2 of 0.9903 and a R2-adjusted of 0.9882. This tells us that 99.03% of the variation is described by the reg6 model, and 98.82% of the variation is described by the indepenedent variables that affect global gross domestic product. Reg7 shows a R2 of 0.9624 and a R2-adjusted of 0.9539. This tells us that 96.24% of the variation is described by the reg7 model, and 95.39% of the variation is described by the indepenedent variables that affect global gross domestic product. For this problem, the R2-adjusted may be the better measure, as across all models most variables seem to be statistically insignificant with total energy consumption being consistently statistically significant. Therefore, we want to ignore those that are insignificant to see the actual robustness of the model.

Elasticity Interpretation

From my calculated, elasticities it seems that global gross domestic product is elastic with respect to the independent variables for most estimated models. This supports my testable hypothesis as we would expect global gross domestic product to go down as things that affect meausres of productivity and efficiency go down. Additionally, it is important note that selecting different units of measurements may affect elasticity. For example changing shipping capacity to thousands of tons will drastically change the elasticity measurement.

Restricted vs Unrestricted model comparison and best model

The F-statistic tests whether a restricted or unrestricted model is better in predicting global gross domestic product. For reg1, it has a f-statistic) of 470.3, and a probability(f-statistic) of 2.34e-22. Based of these values, we must reject the null hythesis, and conclude that reg1 is more statistically significant and a better predictor of violent crime than the naive model. For reg2, it has a f-statistic) of 397.5, and a probability(f-statistic) of 1.10e-17. Based of these values, we must reject the null hythesis, and conclude that reg1 is more statistically significant and a better predictor of violent crime than the naive model. For reg3, it has a f-statistic) of 716.2, and a probability(f-statistic) of 2.35e-24. Based of these values, we must reject the null hythesis, and conclude that reg1 is more statistically significant and a better predictor of violent crime than the naive model. For reg4, it has a f-statistic) of 716.2, and a probability(f-statistic) of 2.35e-24. Based of these values, we must reject the null hythesis, and conclude that reg1 is more statistically significant and a better predictor of violent crime than the naive model. For reg6, it has a f-statistic) of 616.4, and a probability(f-statistic) of 1.07e-23. Based of these values, we must reject the null hythesis, and conclude that reg1 is more statistically significant and a better predictor of violent crime than the naive model. For reg7, it has a f-statistic) of 112.7, and a probability(f-statistic) of 6.43e-15. Based of these values, we must reject the null hythesis, and conclude that reg1 is more statistically significant and a better predictor of violent crime than the naive model.

Based of the AIC score, reg7 is the best model because of it having the minimum score, 120.20, of all models after correcting for autocorrelation. However, reg7 uses all the variables as predictors, and we know that total energy consumption is the most statistically significant variable based of how it has performed, in terms of statsitcal significance, in the other models. Therefore, we conclude that reg7 is the best model however we should conduct further study on the affects of energy consumption and global gross domestic product.

Future model improvements and practical significance

Because so many independent variables ended up being statistically insignificant, we may want to create a model that is more specific by looking at countries with the largest merchant fleets and studying those individual economies and respective GDPs. Then we may factor in other variables such as imports, exports, trade and energy consumption. Additionally, we may want to look at other measures besides GDP that measure efficiency and wellfare as well.

Well done!

Make sure your name is on this notebook at the top and on the file.
Please submit this notebook as a PDF file. Nothing else will be accepted.