Midterm

UNIVERSITY OF
SAN FRANCISCO

Econometrics Midterm

Prof. Jesse Anttila-Hughes

Student: Shiv Gargé

Student code: 20702236

Course: Econ 620-03 Graduate Econometrics

Spring 2023

Declaration

I hereby certify that this midterm constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another. I declare that the midterm describes original work that has not previously been presented for the qualification of any other degree of any institution.

Shiv Gargé

Contents

1. Short Answers

1.1 Question 1.A

The ordinary least squares (OLS) regression estimator is a linear estimator due to its assumption that the relationship between the dependent and independent variables is linear. This assumption implies that there exists a constant rate of change in the dependent variable for each unit increase in an independent variable, which is represented by the regression coefficient. In the OLS model, it is assumed that adding any number of independent variables is additive and independent of the other independent variables in the model. Consequently, the coefficients of a regression output equation represent the implied change in the dependent variable (y-hat) for a single unit increase in an independent variable while holding all other independent variables constant. For example, Equation alpha shows a regression equation:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$\hat{\alpha} = 3.76 + 0.735 X_1 + 6.48 X_2 + 1.74 X_3$$

In equation *alpha*, the independent variables are X_1, X_2, X_3 which have coefficients 0.735, 6.48 and 1.74 respectively. The additive assumption means that the effect of one independent variable is independent of the other independent variables. In equation *alpha*, this translates to the impact of a one-unit increase in X_1 on the dependent variable is the same regardless of the value of X_2 or X_3. Secondly, the assumption of independence means that the independent variables are not correlated with each other. This means that the relationship between the dependent variable and each independent variable can be analyzed separately without affecting the relationship between the dependent variable and the other independent variables.

1.2 Question 1.B

Due to the lack of clarity on this question, I will present two answers.

Firstly is a model-based answer. In the case that the dependent variable is binary such as A or B, the approach would be to use logistic regression (logit). This model is used to identify the probability that the dependent variable changes from A to B:

$$P(Y = B|X) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k\right)\right)}$$

The general form of a logit regression uses a sigmoid function which is a non-linear function. This violates the assumption of linearity; hence it cannot be used to estimate the parameters.

Secondly, exponential regression is used to describe an underlying exponential function, with the general form being as follows:

$$Y = \beta_0 e^{\beta_1 X_1}$$

Again, OLS cannot be used for such a model since it also violates the linearity assumption.

This follows the second interpretation of the question. The first example of a model that OLS cannot be used for is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2^3 + \varepsilon$$

In this example, the terms X_1^2 and X_2^3 make this equation non-liner due to the introduction of a polynomial relationship between the independent variables and the dependent variable. This violates the assumption of linearity for OLS to be used effectively.

The second example is as follows:

$$Y = \beta_0 + \beta_1 e^{\beta_2 X_1} + \epsilon$$

In this example, the term $e^{\beta_2 X_1}$ introduces an exponential relationship between the independent and dependent variable, which yet again violates the assumption of linearity hence rendering OLS ineffective.

1.3 Question 2.A

There are five main assumption needed for the best linear Unbiased Estimate (BLUE) which are as follows:

1. Linearity - The model must be linear in the parameters, which results in being able to represent the relationship between dependent and independent variables as a linear function of the underlying parameters used.
2. Homoscedacticity - This refers to the error term having a constant variance. If the error term has inconsistent variance in the error term, the error term is heteroscedastic.
3. Independence - The value of the error term from one observation should not affect or be related to the value of the independednt variable from the same observation.
4. Sufficiency - This is the assumption that there is an adequate number of observations or data to estimate the parameters with some degree of accuracy.
5. Exogeneity - The independednt variable must be exogenous, which implies that they are not correlated to the error term.

1.4 Question 2.B

i) In the case that homoscedasticity is violated, the OLS estimator would still be unbiased, but the change occurs in the standard errors, which may be wrong. This can cause issues such as incorrect confidence intervals and invalid hypothesis testing.

OLS assigns equal weights to each observation, so if the errors have different variances, some observations will count toward the estimator more than others.

ii) If there is a study on health outcomes and income. With a large sample, some of whom have very low incomes while others have very high incomes. In this case, it is possible that the variability in the dependent variable (health outcomes) may increase or decrease as income increases. For example, individuals with low incomes may have a wider range of health outcomes than those with high incomes, resulting in heteroscedasticity.

1.5 Question 3.A

The three components of variance in the OLS model are as follows:

1. Sampling - This is the variation that occurs from the sampling variability. It shows that different samples will produce different estimates of the regression coefficients. This is considered to be 'good' variance which can result in wider confidence intervals and less accurate statistical inference.
2. Specification - This is the variation that arises from incorrectly specifying the regression model. It shows that the estimated regression coefficients may be biased due to omitted variables or measurement errors. This can lead to biased estimates of the regression coefficients. This is considered to be 'bad' variance which can result in incorrect statistical inference.
3. Error Term - This component represents the variation in the dependent variable (y) that is not explained by the independent variables (x). It is considered as 'good' variation because it is a random variation that is not related to the independent variables. which can result in wider confidence intervals and less accurate statistical inference.

1.6 Question 3.B

As the number of observations (N) approaches infinity, The sampling variance approaches zero. This results in the estimator becoming more accurate along with the estimated coefficients increasing in accuracy. For the error term variance, It may not approach zero as N approaches infinity. This can be caused by the data generation process and if it is inherently noisy. For specification, N approaching infinity has no effect on the variation due to the underlaying issues of model specification and error in data generation. However, It is important to note that even if these components do not approach zero with increasing sample size, their influence on the estimator will become relatively smaller as the sample size increases, allowing the 'good' variation component to dominate and improving the accuracy and precision of the estimator.

1.7 Question 4.A

Univariate OLS models with a highly statistically significant coefficient on X and a high R-squared have similarities and differences. Both models suggest a strong relationship between the independent variable X and the dependent variable Y and indicate that X

is a good predictor of Y. However, a highly significant coefficient on X indicates that X is important in explaining the variation in Y, while a high R-squared indicates that a large proportion of the variation in Y can be explained by X alone. In contrast, a significant coefficient on X may be obtained even when the model explains only a small proportion of the variation in Y or includes other variables that are not statistically significant. On the other hand, a high R-squared suggests a good fit of the model to the data, but it does not necessarily imply statistical significance of individual coefficients on the independent variables.

1.8 Question 4.B

A univariate OLS model may exhibit a high coefficient of determination (R-squared) despite a low t-statistic. This implies that the model provides a good fit to the data, but the relationship between the independent and dependent variables may be weak or non-significant. Conversely, the model may show a high t-statistic but a low R-squared, indicating a significant relationship between the variables but poor explanatory power of the model. This may result from a failure to account for other important variables or model misspecification.

2. Long Answers

2.1 Question 5.A

The authors' main statistical model tested whether looking at selfies on social media impacts people's self-esteem. The model included variables such as gender, age, frequency of selfie and groupie viewing, as well as how often people post selfies and groupies themselves.

The coefficient on "Selfie Viewing Frequency" in the model reflects how much the frequency of looking at selfies on social media is associated with people's self-esteem, after taking into account other factors in the model. If this coefficient is negative and statistically significant, it means that viewing selfies more often is linked to lower self-esteem.

$$\begin{aligned} \text{Self-Esteem} = \beta_0 &+ \beta_1 \text{Gender} + \beta_2 \text{Age} \\ &+ \beta_3 \text{Frequency of Selfie Viewing} + \beta_4 \text{Frequency of Groupie Viewing} \\ &+ \beta_5 \text{Selfie Posting Frequency} + \beta_6 \text{Groupie Posting Frequency} + \epsilon \end{aligned} \tag{1}$$

where β_2 is the intercept, Gender is a binary variable representing male or female, Age is a continuous variable representing age in years, Frequency of Selfie Viewing is a continuous variable representing how often the participant views selfies, Frequency of Groupie Viewing is a continuous variable representing how often the participant views groupies, Selfie Posting Frequency is a continuous variable representing how often the participant posts selfies, Groupie Posting Frequency is a continuous variable representing how often the participant posts groupies, and epsilon is the error term.

The coefficient on Selfie Viewing Frequency β_3 represents the effect of the frequency of selfie viewing on self-esteem, controlling for other variables in the model. A negative coefficient on this variable would indicate that as the frequency of selfie viewing increases, self-esteem decreases, after controlling for the effects of the other variables in the model.

Therefore, if the coefficient on Selfie Viewing Frequency is negative and statistically significant, it would suggest that frequent exposure to selfies is associated with lower levels of self-esteem.

2.2 Question 5.B

The researchers in the article assumed that the errors in their model had a mean of zero, which means that the errors were not biased in any systematic way. This is important because if the errors are biased, it could lead to inaccurate estimates of how the independent variables affect the dependent variable.

However, the research design of the study may have violated this assumption. The study only collected data from participants at a single point in time, which means that the researchers couldn't measure how the variables changed over time. For example, they couldn't measure if the frequency of selfie viewing and posting changed over time, or if changes in self-esteem led to changes in the way individuals take their selfies. This limits the researchers' ability to make causal claims about the relationship between the variables.

Additionally, the variables in the study were all self-reported by the participants. While self-reported measures can be useful, they are also subject to social desirability bias. This means that participants may report their behavior or feelings in a way that they believe is socially acceptable, rather than being completely truthful. This can lead to measurement errors in the model and make it more difficult to accurately estimate the true relationship between the variables. Additionally, I have looked at the issue of self-reporting in my undergraduate thesis. There is strong literature supporting the use of experimental games which have been shown to reduce the instances of overreporting [1].

If the assumption of zero-mean errors is violated, the results may be biased. For example, if the errors are correlated with unmeasured variables, like personality traits or other psychological factors, this could lead to biased estimates of the relationship between the variables. This is important because it could lead to incorrect conclusions about the relationship between the variables and limit the ability to apply the results to other groups of people.

Adding more variables to the model may help to address the issue of omitted variable bias and measurement error, but it may not completely solve the problem. Using data collected over a longer period of time or conducting experiments may also help to provide stronger evidence for causal relationships between the variables.

2.3 Question 5.C

An additional social factor that may be correlated with self-esteem and is not included as one of the controls in Model 2 is social support. Social support, such as having close relationships with friends or family, has been found to be positively related to self-esteem.

$$\begin{aligned} Self - Esteem =& \beta_0 + \beta_1 Gender + \beta_2 Age + \beta_3 Freq_{selfie} \\ &+ \beta_4 Freq_{groupie} + \beta_5 Post_{selfie} + \beta_6 Post_{groupie} + \beta_7 Social\ Support + \epsilon \end{aligned}$$

Including social support as a control variable in the regression model may help to address the potential issue of omitted variable bias. Omitted variable bias occurs when there is a variable that is correlated with both the independent and dependent variables, and that is not included in the model as a control. This can lead to biased estimates of the coefficients on the included variables. By including social support as a control, we may be able to reduce the influence of other unmeasured variables that are related to both selfie behavior and self-esteem.

Social support as a control variable. Control variables are used to isolate the relationship between the independent variable(s) and the dependent variable, by holding constant the effects of other variables that may be related to the dependent variable. The model can give us more insight into how it might affect the relationship between taking selfies and self-esteem. If social support plays a moderating role, it could mean that people with high social support may experience different effects from taking selfies on their self-esteem than those with low social support. This information could help us design interventions or programs to improve self-esteem by taking into account social support levels.

Adding social support as a control variable to the model may not be enough to fully address the issue of bias caused by variables that were not measured in the study but are related to both taking selfies and self-esteem. Moreover, the data used in the study is collected at a single point in time, which means it is not possible to establish a clear cause-and-effect relationship between the variables. As a result, we need to be cautious in drawing conclusions about the relationship between taking selfies and self-esteem, and consider the possibility that there may be other factors that we have not accounted for.

2.4 Question 5.D

A possible research design that could better determine the causal relationship between viewing selfies and reduced self-esteem is a randomized controlled trial (RCT). In this design, participants would be randomly assigned to either a treatment or a control group. The treatment group would be exposed to selfies, while the control group would not be exposed.

To ensure that the study results are representative, participants from different age groups, genders, and ethnicities should be included in the study. The sample size should be large enough to ensure adequate statistical power. Additionally, participants

should be screened for any pre-existing mental health conditions or history of low self-esteem.

To make sure that our findings are reliable and not influenced by any biases, we need to take some steps to ensure that our study is conducted properly. We will measure participants' self-esteem before and after they view selfies to see if there is any change. It's important that participants don't know which group they are in (treatment or control), so that their behavior is not influenced by this knowledge. Similarly, the experimenters who collect the data should also not know which group each participant is in to prevent any potential bias.

One way to analyze the data collected in this research design is by using a method called difference-in-differences. This method can help estimate the impact of viewing selfies on self-esteem, while accounting for any differences between the treatment and control groups before the study began. Using this research design allows us to better understand the cause-and-effect relationship between viewing selfies and self-esteem. This approach can help control for other factors that might be influencing self-esteem, as well as addressing concerns about the accuracy of our conclusions.

3. Jupyter Notebook Continues at end of Document

3.1 Reason for Python over Stata

I was having significant issues running the questions on Stata. It is a reflection of my lack of experience with Stata which is something I need to work on.

Bibliography

[1] Isabel Thielmann, Robert Böhm, Marion Ott, and Benjamin E. Hilbig. Economic Games: An Introduction and Guide for Research. *Collabra: Psychology*, 7(1), 02 2021. ISSN 2474-7394. doi: 10.1525/collabra.19004. URL https://doi.org/10.1525/collabra.19004. 19004.

In [13]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
from sklearn.impute import KNNImputer
```

In [2]:
```
#PART A
#Load the data
df = pd.read_stata(r"C:\Users\shivg\OneDrive\Desktop\population_gdp_panel.dta")

# Convert columns to float
df['gdp'] = pd.to_numeric(df['gdp'], errors='coerce')
df['population'] = pd.to_numeric(df['population'], errors='coerce')
df['year'] = pd.to_numeric(df['year'], errors='coerce')
```

In [3]:
```
# Create GDP per capita variable
df["GDPpc"] = df["gdp"] / df["population"]

# Calculate average GDPpc for each country
grouped = df.groupby("country")["GDPpc"].mean().reset_index()

# Create income groups based on average GDPpc
lower_income = grouped[grouped["GDPpc"] < 3000]["country"]
middle_income = grouped[(grouped["GDPpc"] >= 3000) & (grouped["GDPpc"] < 10000)]["countr
upper_income = grouped[grouped["GDPpc"] >= 10000]["country"]

# Add income group variable to original data
df["income_group"] = ""
df.loc[df["country"].isin(lower_income), "income_group"] = "lower"
df.loc[df["country"].isin(middle_income), "income_group"] = "middle"
df.loc[df["country"].isin(upper_income), "income_group"] = "upper"

# Calculate summary statistics for each income group
summary = df.groupby("income_group").agg({"gdp": ["mean", "std", "min", "max"],
                                          "population": ["mean", "std", "min", "max"],
                                          "GDPpc": ["mean", "std", "min", "max"]})
```

In [4]:
```
df
```

Out[4]:

	country	year	population	gdp	GDPpc	income_group
0	Afghanistan	1970	11126123.0	1.072092e+10	963.580868	lower
1	Afghanistan	1971	11417825.0	1.069136e+10	936.374543	lower
2	Afghanistan	1972	11721940.0	8.937847e+09	762.488731	lower
3	Afghanistan	1973	12027822.0	9.196673e+09	764.616694	lower
4	Afghanistan	1974	12321541.0	9.698170e+09	787.090669	lower
...	...	...	...	...	...	...
10555	Zimbabwe	2013	15054506.0	1.418193e+10	942.038655	lower
10556	Zimbabwe	2014	15411675.0	1.448359e+10	939.780273	lower
10557	Zimbabwe	2015	15777451.0	1.472830e+10	933.503264	lower
10558	Zimbabwe	2016	16150362.0	1.481899e+10	917.563715	lower
10559	Zimbabwe	2017	16529904.0	1.526452e+10	923.448592	lower

10560 rows × 6 columns

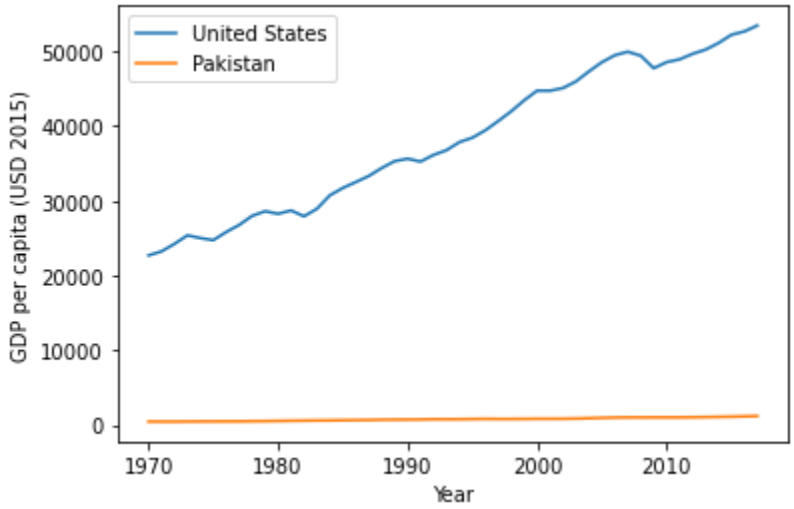
In [5]:
```
#PART B
# two countries from different income groups
country1 = 'United States'
country2 = 'Pakistan'

# Subset the original dataframe for these two countries
df_country1 = df[df['country'] == country1]
df_country2 = df[df['country'] == country2]

# Plot the line charts
plt.plot(df_country1['year'], df_country1['GDPpc'], label=country1)
plt.plot(df_country2['year'], df_country2['GDPpc'], label=country2)

# Add axis labels and a legend
plt.xlabel('Year')
plt.ylabel('GDP per capita (USD 2015)')
plt.legend()

# Show the plot
plt.show()
```

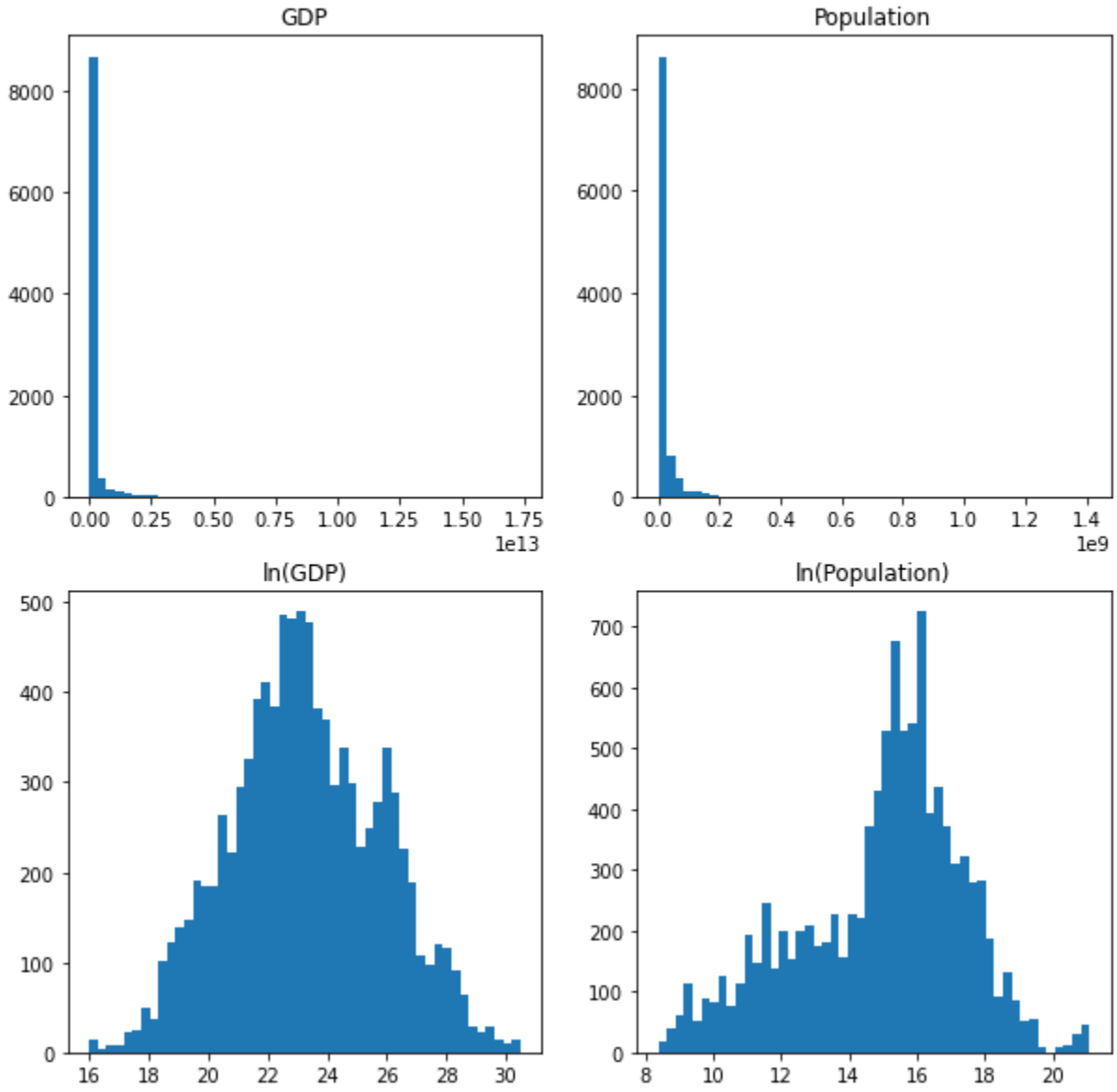


While there is no direct indication for Pakistan, you can clearly see a significant dip in the the US GDP per capita in 2008. This is the year of the Great Recession, which was a global economic crisis that started in the US and spread to other countries. The US GDP per capita has since recovered. Pakistan's GDP per capita has been relatively stable over the years with a slight increase in the last few years. This shows lack of economic growth in Pakistan. at the enf of 2020, the US had a GDP per capita of 60,000 USD while Pakistan had a GDP per capita of 1,800 USD.

In [9]:
```
#PART C
# Create histograms of GDP, population, ln(GDP), and ln(population)
fig, axs = plt.subplots(2, 2, figsize=(10, 10))
axs[0, 0].hist(df['gdp'], bins=50)
axs[0, 0].set_title('GDP')
axs[0, 1].hist(df['population'], bins=50)
axs[0, 1].set_title('Population')
axs[1, 0].hist(np.log(df['gdp']), bins=50)
axs[1, 0].set_title('ln(GDP)')
axs[1, 1].hist(np.log(df['population']), bins=50)
axs[1, 1].set_title('ln(Population)')

df['lnPopulation'] = np.log(df['population'])
df['lnGDP'] = np.log(df['gdp'])
```

```
# Show the plot
plt.show()
```



In [17]:
```
# Count the number of missing values in each column of the DataFrame
missing_values_count = df.isnull().sum()

# Print the total number of missing values in the DataFrame
print("Total number of missing values in the DataFrame:", missing_values_count.sum())
```

```
Total number of missing values in the DataFrame: 0
```

In [27]:
```
#PART D
#cross sectional regression of lnGDP on lnPop for year 2000
sub_df = df[df["year"] == 2000]
X = sub_df[["lnPopulation"]]
X = sm.add_constant(X)
y = sub_df["lnGDP"]
model = sm.OLS(y, X).fit()
print(model.summary())
r_squared = model.rsquared
print("R-squared:", r_squared)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  lnGDP   R-squared:                       0.603
Model:                            OLS   Adj. R-squared:                  0.601
Method:                 Least Squares   F-statistic:                     316.0
Date:                Fri, 31 Mar 2023   Prob (F-statistic):           1.31e-43
Time:                        20:54:03   Log-Likelihood:                -389.30
No. Observations:                 210   AIC:                             782.6
Df Residuals:                     208   BIC:                             789.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          11.4495      0.682     16.786      0.000      10.105      12.794
lnPopulation    0.7951      0.045     17.777      0.000       0.707       0.883
==============================================================================
Omnibus:                       19.314   Durbin-Watson:                   1.779
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                7.262
Skew:                           0.164   Prob(JB):                       0.0265
Kurtosis:                       2.150   Cond. No.                         97.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
R-squared: 0.6030619819300742
```

The OLS regression results provide information about the relationship between the dependent variable (lnGDP in this case) and the independent variable (lnPopulation in this case).

The R-squared value of 0.603 indicates that approximately 60.3% of the variation in lnGDP can be explained by lnPopulation. This means that the model explains a moderate amount of the variability in the dependent variable, and there may be other factors that contribute to the variation in lnGDP that are not accounted for by the model.

The coefficient of lnPopulation is 0.7951, which indicates that for each 1% increase in lnPopulation, lnGDP is predicted to increase by approximately 0.7951%.

The standard error of the coefficient is given by std err, which measures the precision of the estimate of the coefficient. In this case, the standard error of lnPopulation is 0.045.

The t-value of 17.777 and the associated p-value of 0.000 indicate that the coefficient of lnPopulation is statistically significant at the 5% level. This means that we can reject the null hypothesis that there is no relationship between lnPopulation and lnGDP.

The intercept of 11.4495 represents the predicted value of lnGDP when lnPopulation is equal to zero. However, this value is not meaningful in this case since it is not possible for population to be zero.

In [28]:
```python
#PART E
#regression of lnGDP on lnPop and year
X = df[["lnPopulation", "year"]]
X = sm.add_constant(X)
y = df["lnGDP"]
model = sm.OLS(y, X).fit()
print(model.summary())
r_squared = model.rsquared
print("R-squared:", r_squared)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  lnGDP   R-squared:                       0.649
Model:                            OLS   Adj. R-squared:                  0.649
Method:                 Least Squares   F-statistic:                     8910.
Date:                Fri, 31 Mar 2023   Prob (F-statistic):               0.00
Time:                        20:54:14   Log-Likelihood:                -17664.
No. Observations:                9658   AIC:                         3.533e+04
Df Residuals:                    9655   BIC:                         3.536e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -23.8938      2.223    -10.748      0.000     -28.252     -19.536
lnPopulation   0.8186      0.006    130.448      0.000       0.806       0.831
year           0.0175      0.001     15.669      0.000       0.015       0.020
==============================================================================
Omnibus:                     1201.518   Durbin-Watson:                   0.041
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              339.908
Skew:                           0.141   Prob(JB):                     1.55e-74
Kurtosis:                       2.125   Cond. No.                     2.89e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
[2] The condition number is large, 2.89e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
R-squared: 0.648596092050278
```

the model has an R-squared value of 0.649, indicating that approximately 65% of the variation in the dependent variable (lnGDP) can be explained by the independent variables (lnPopulation and year). The adjusted R-squared value is the same, indicating that the addition of the independent variables did not improve the model's fit.

The coefficient for lnPopulation is 0.8186, indicating that for each 1% increase in the natural logarithm of the population, there is an estimated 0.8186% increase in lnGDP, all other variables being held constant.

The coefficient for year is 0.0175, indicating that for each additional year, there is an estimated 0.0175 increase in lnGDP, all other variables being held constant.

The constant term (-23.8938) is the estimated value of lnGDP when lnPopulation is zero and year is zero.

In [29]:

```python
#PART F
#replicate regression for low and high income subsamples
low_income_df = df[df["income_group"] == "lower"]
X = low_income_df[["lnPopulation", "year"]]
X = sm.add_constant(X)
y = low_income_df["lnGDP"]
model = sm.OLS(y, X).fit()
print(model.summary())

high_income_df = df[df["income_group"] == "upper"]
X = high_income_df[["lnPopulation", "year"]]
X = sm.add_constant(X)
y = high_income_df["lnGDP"]
model = sm.OLS(y, X).fit()
print(model.summary())
r_squared = model.rsquared
print("R-squared:", r_squared)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  lnGDP   R-squared:                       0.859
Model:                            OLS   Adj. R-squared:                  0.859
Method:                 Least Squares   F-statistic:                 1.286e+04
Date:                Fri, 31 Mar 2023   Prob (F-statistic):               0.00
Time:                        20:54:31   Log-Likelihood:                -4781.9
No. Observations:                4233   AIC:                             9570.
Df Residuals:                    4230   BIC:                             9589.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -21.2671      1.674    -12.706      0.000     -24.549     -17.986
lnPopulation   0.8712      0.006    155.192      0.000       0.860       0.882
year           0.0151      0.001     17.913      0.000       0.013       0.017
==============================================================================
Omnibus:                       92.565   Durbin-Watson:                   0.050
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               97.069
Skew:                          -0.362   Prob(JB):                     8.35e-22
Kurtosis:                       2.837   Cond. No.                     2.90e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
[2] The condition number is large, 2.9e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                  lnGDP   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                 2.273e+04
Date:                Fri, 31 Mar 2023   Prob (F-statistic):               0.00
Time:                        20:54:31   Log-Likelihood:                -3304.1
No. Observations:                3135   AIC:                             6614.
Df Residuals:                    3132   BIC:                             6632.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -18.4367      1.791    -10.292      0.000     -21.949     -14.924
lnPopulation   0.9890      0.005    210.847      0.000       0.980       0.998
year           0.0144      0.001     16.010      0.000       0.013       0.016
==============================================================================
Omnibus:                       14.636   Durbin-Watson:                   0.074
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               15.473
Skew:                          -0.133   Prob(JB):                     0.000437
Kurtosis:                       3.218   Cond. No.                     2.88e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
[2] The condition number is large, 2.88e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
R-squared: 0.9355355348876548
```

For both OLS regression models, the R-squared values are high, indicating that a large proportion of the variance in the dependent variable (lnGDP) is explained by the independent variables (lnPopulation and year). In the first model, R-squared is 0.859, while in the second model, R-squared is 0.936. The adjusted

R-squared values are also high, suggesting that the independent variables in the models are good predictors of lnGDP.

In both models, the coefficient for lnPopulation is positive and statistically significant, with a p-value of 0.000, indicating that as the natural log of the population increases, the natural log of GDP also increases.

Similarly, the coefficient for year is also positive and statistically significant, with a p-value of 0.000, implying that as the year increases, the natural log of GDP also increases.

The standard errors for the coefficients in both models are also reported, which can be used to calculate the confidence interval for each coefficient. For example, the 95% confidence interval for the coefficient of lnPopulation in the first model is (0.786, 0.851), while the 95% confidence interval for the coefficient of lnPopulation in the second model is (0.927, 0.967).

In [23]:
```
#Part G
#partial out lnGDP and lnPop by year for the whole sample
X = df[["year"]]
X = sm.add_constant(X)
```

In [24]:
```
# regression of lnGDP on year
y = df["lnGDP"]
model1 = sm.OLS(y, X).fit()
df["lnGDP_resid"] = model1.resid
```

In [25]:
```
# regression of lnPop on year
y = df["lnPopulation"]
model2 = sm.OLS(y, X).fit()
df["lnPopulation_resid"] = model2.resid
```

In [30]:
```
# regression of residual lnGDP on residual lnPop
X = df[["lnPopulation_resid"]]
X = sm.add_constant(X)
y = df["lnGDP_resid"]
model3 = sm.OLS(y, X).fit()
print(model3.summary())
r_squared = model.rsquared
print("R-squared:", r_squared)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            lnGDP_resid   R-squared:                       0.638
Model:                            OLS   Adj. R-squared:                  0.638
Method:                 Least Squares   F-statistic:                 1.702e+04
Date:                Fri, 31 Mar 2023   Prob (F-statistic):               0.00
Time:                        20:54:36   Log-Likelihood:                -17664.
No. Observations:                9658   AIC:                         3.533e+04
Df Residuals:                    9656   BIC:                         3.535e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const               1.068e-13      0.015   6.96e-12      1.000      -0.030       0.030
lnPopulation_resid     0.8186      0.006    130.454      0.000       0.806       0.831
==============================================================================
Omnibus:                     1201.518   Durbin-Watson:                   0.041
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              339.908
Skew:                           0.141   Prob(JB):                     1.55e-74
Kurtosis:                       2.125   Cond. No.                         2.44
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
R-squared: 0.9355355348876548
```

The coefficient for lnPopulation_resid is statistically significant (p-value<0.05), indicating a positive relationship between lnGDP_resid and lnPopulation_resid. The R-squared value of 0.638 indicates that the independent variable explains 63.8% of the variance in the dependent variable.

The regression that "partially outs" lnGDP and lnPopulation by year for the whole sample by first separately regressing lnGDP and lnPopulation on year, predicting residuals from each of those regressions, and then running the regression of residual Ys on residual Xs is essentially a regression of the two variables after removing the effect of year. This regression is also referred to as a "fixed effects" regression, where the individual-specific effects (in this case, year-specific effects) are controlled for.

The regression in part e is a simple linear regression of lnGDP_resid on lnPopulation_resid, without controlling for the year-specific effects. Therefore, the regression in part e does not control for any time-specific factors that may affect the relationship between lnGDP_resid and lnPopulation_resid, while the fixed effects regression in the second part does control for time-specific factors.

In [ ]: