Lab 4

Sevastian Sanchez

Naive OLS & First Difference Analysis

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

```
# grab the data from online, Excel is fine
url = 'https://www.qogdata.pol.gu.se/data/qog_bas_ts_jan24.xlsx'
df = pd.read_excel(url)
```

```
df.head()
```

	ccode	cname	year	ccode_qog	cname_qog	ccodealp	ccodecow	version	cname_year	ccodealp_year	...	wdi_trade
0	4	Afghanistan	1946	4	Afghanistan	AFG	700.0	QoGBasTSjan24	Afghanistan 1946	AFG46	...	NaN
1	4	Afghanistan	1947	4	Afghanistan	AFG	700.0	QoGBasTSjan24	Afghanistan 1947	AFG47	...	NaN
2	4	Afghanistan	1948	4	Afghanistan	AFG	700.0	QoGBasTSjan24	Afghanistan 1948	AFG48	...	NaN
3	4	Afghanistan	1949	4	Afghanistan	AFG	700.0	QoGBasTSjan24	Afghanistan 1949	AFG49	...	NaN
4	4	Afghanistan	1950	4	Afghanistan	AFG	700.0	QoGBasTSjan24	Afghanistan 1950	AFG50	...	NaN

5 rows × 251 columns

1-- Run a naive OLS regression on your time series data. Tell me how you expect your Xs to affect your Y and why. Interpret your results.

I am going to predict CO2 emissions of countries.

```
df[['wdi_co2']].describe()
```

	wdi_co2
count	5803.000000
mean	4.226968
std	5.456033
min	0.000000
25%	0.587717
50%	2.261268
75%	6.162758
max	47.656962

The next variable I will be looking at is how much of a country's population is urban (%). The wdi_popurb indicator, which measures urban population as a percentage of total population, can theoretically range from 0% to over 100%, depending on the extent of a country's population residing in urban regions.

0% indicates no urban population, meaning the country has no urban population relative to its total population (i.e. rural being the inverse of urban in this case) 100% indicates that the proportion of urban population of a country is equal to to its total population (i.e. 100% of a given country's total population is urban)

```
df[['wdi_popurb']].describe()
```

	wdi_popurb
count	10490.000000
mean	50.709539
std	24.858128
min	2.193000
25%	29.807750
50%	49.938000
75%	71.465000
max	100.000000

Another X variable I will be using to predict CO2 emissions is GDP per capita. To better standardize GDP, I will be taking the log GDP.

Let's look at the description of this variable.

```
df['log_gle_cgdpc']=np.log(df['gle_cgdpc'])
df[['log_gle_cgdpc']].describe()
```

	log_gle_cgdpc
count	9478.000000
mean	7.656398
std	1.481882
min	3.983599
25%	6.538187
50%	7.590481
75%	8.744678
max	11.943892

I am now going to predict CO2 emissions as a function of urbanization (% of urban pop) and GDP per capita. I expect countries with greater proportions of their population residing in urban areas and higher GDP per capita to have higher rates of CO2 emissions per metric ton(MT).

```
co2_1 = smf.ols(formula = 'wdi_co2 ~ wdi_popurb + log_gle_cgdpc', data = df).fit()
print (co2_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                wdi_co2   R-squared:                       0.522
Model:                            OLS   Adj. R-squared:                  0.522
Method:                 Least Squares   F-statistic:                     2232.
Date:                Wed, 20 Nov 2024   Prob (F-statistic):               0.00
Time:                        01:47:45   Log-Likelihood:                -11354.
No. Observations:                4083   AIC:                         2.271e+04
Df Residuals:                    4080   BIC:                         2.273e+04
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept       -18.2247      0.446    -40.818      0.000     -19.100     -17.349
wdi_popurb        0.0614      0.004     16.711      0.000       0.054       0.069
log_gle_cgdpc     2.2939      0.066     34.598      0.000       2.164       2.424
==============================================================================
Omnibus:                     2937.442   Durbin-Watson:                   0.107
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            63852.323
Skew:                           3.200   Prob(JB):                         0.00
Kurtosis:                      21.286   Cond. No.                         431.
==============================================================================

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

Net of urbanization, a one-unit increase in Log GDP per capita (log_gle_cgdpc), say, going from 0.5 to 1.5, is associated with an approximate 2.29 point increase in CO2 emissions per MT. Likewise, net of log gdp per capita, a one-unit increase in urban population (wdi_popurb) is associated with an approximate 0.06 point increase in CO2 emissions per MT. Both relationships are statistically significant, with p-values less than 0.05.

Overall, those two variables allow us to predict CO2 emissions with 52% more accuracy compared to merely guessing the average CO2 MT.

2-- Run a first differences regression on the same model in Question 1. Interpret your results. Do you draw a different conclusion than in Question 1? Explain.

Now, let's look at the first differences of these relationship within the countries over time. Will it continue to be the case that higher urbanization and GDP are associated with higher CO2 Emissions?

```
!pip install linearmodels

from linearmodels.panel import FirstDifferenceOLS
```

Show hidden output

```
columns = ['ccode', 'year', 'wdi_co2', 'wdi_popurb', 'log_gle_cgdpc']
df1 = df[columns].dropna()


# Set the MultiIndex for panel data
df1 = df1.set_index(['ccode', 'year'])

# Define the dependent and independent variables
y = df1['wdi_co2']
X = df1[['wdi_popurb', 'log_gle_cgdpc']]

# Fit the first-differenced panel data model
fdmodel = FirstDifferenceOLS(y, X)
results = fdmodel.fit(cov_type='clustered', cluster_entity=True)

print(results)
```

```
                     FirstDifferenceOLS Estimation Summary
================================================================================
Dep. Variable:                wdi_co2   R-squared:                        0.0359
Estimator:         FirstDifferenceOLS   R-squared (Between):              0.2803
No. Observations:                3890   R-squared (Within):               0.0691
Date:                Wed, Nov 20 2024   R-squared (Overall):              0.2701
Time:                        02:20:45   Log-likelihood                   -2960.8
Cov. Estimator:             Clustered
                                        F-statistic:                      72.384
Entities:                         193   P-value                           0.0000
Avg Obs:                       21.155   Distribution:                  F(2,3888)
Min Obs:                       1.0000
Max Obs:                       22.000   F-statistic (robust):             9.0258
                                        P-value                           0.0001
Time periods:                      22   Distribution:                  F(2,3888)
Avg Obs:                       185.59
Min Obs:                       163.00
Max Obs:                       191.00

                              Parameter Estimates
=================================================================================
               Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
---------------------------------------------------------------------------------
wdi_popurb       -0.0063     0.0277    -0.2282     0.8195     -0.0606      0.0480
log_gle_cgdpc     0.9628     0.4675     2.0593     0.0395      0.0461      1.8794
=================================================================================
```

Net of urbanization, a one-unit increase in Log GDP per capita (log_gle_cgdpc)is associated with an approximate 96.28% change in CO2 emissions per MT , and with statistical significance at a p-value less than 0.05 (~0.04).

However, net of log gdp per capita, a one-unit increase in urban population (wdi_popurb) is associated with an approximate -0.63% change in CO2 emissions per MT from one period to the next, and with no statistical significance having a p-value greater than 0.05 (~0.82).

Note that the coefficients are smaller in magnitude compared to the levels model, which is expected since we're now looking at changes in CO2 emissions from one period to the next, rather than the overall levels of CO2.

Overall, we still see that Log GDP has a substantively significant and positive impact on CO2 emission MT growth. However, urbanization has a very small negative (i.e. direction) impact on CO2 emission MT growth, and no statistical significance. Log GDP appears much more related to CO2 Emission growth than urbanization.

Just so we can see how the first differences are distributed, take a look at them here.

```
# Group by 'ccode' and calculate the first differences for the relevant columns

df1['wdi_co2_diff'] = df1.groupby('ccode')['wdi_co2'].diff()
df1['wdi_popurb_diff'] = df1.groupby('ccode')['wdi_popurb'].diff()
df1['log_gle_cgdpc_diff'] = df1.groupby('ccode')['log_gle_cgdpc'].diff()


df1[['wdi_co2_diff', 'wdi_popurb_diff', 'log_gle_cgdpc_diff']].describe()
```

	wdi_co2_diff	wdi_popurb_diff	log_gle_cgdpc_diff
count	3890.000000	3890.000000	3890.000000
mean	0.011040	0.302722	0.037469
std	0.527500	0.410814	0.097793
min	-11.550203	-3.028000	-1.052427
25%	-0.045765	0.046000	0.000477
50%	0.007215	0.249000	0.038306
75%	0.095372	0.492750	0.077877
max	8.379716	3.359000	1.129050