# 7ECON015W.2 Applied Economics 2024-2025
# Computer practical 5-7

This exercise is designed to help you to build your data sorting skill in STATA. There are many publicly accessible datasets, but you need to be careful with the data source and its quality. Of course, at the county-level, the WB, IMF, OECD, etc., are all the good sources. However, there are not many good data resources at the sectoral level. The EU KLEMS is a high quality, trustable, and very well-known dataset at the sector-country-year level. It covers 27 EU countries, and also the data for the UK, US, and Japan. In this practice, we mainly focus on using EU KLEMS data to analyse the impact of variable and fixed inputs on labour productivity and TFP across 30 countries (*assuming everything else is constant*).

1. Visit Luiss Lab of European Economics website: https://euklems-intanprod-llee.luiss.it/download/ and then click "CSV" under "National Account" and "All Countries from 1995"

2. Now you should have your dataset downloaded. Be clear where is the file. Use the following command to load the CSV in STATA (you should open a new do-file editor and again write the code within the do file):

```
clear all
cap log close
set more off
set matsize 800
set maxvar 20000
cd "E:\MSc Applied Economic Module 2025\Exercise\STATA Practical 5-6"
     ^^^^^^^^^^^^^^^(you need to change the routing)^^^^^^^^^^^^^^
```

```
import delimited using "E:\MSc Applied Economic Module 2025\Exercise\STATA Practical 5-6\national_accounts.csv", varnames(1)
```

3. Browse the dataset. Are all variable highlighted as red? It means that all data is coded in string. This is really bad as string data cannot be run in the regression. We need to transform part of them into numeric. Use command:

```
drop v1

destring year - va_q, force replace

rename geo_name country
rename nace_r2_name sector
```

4. Browse the dataset. The dataset contains a variable that indicates the "sector", labelled as "sector" (we just renamed it from step 3). There is also a variable call "country" (again we just renamed it from step 3). These two variables are in red still. Again, we need to code them in "numeric" but we also want to keep these string

information in the dataset as otherwise we would not know which sector is connected to the number. Use command:

```
encode sector, gen(n_sector)
encode country, gen(n_country)
```

5. Now check if the code works. Use command:

```
ed sector n_sector
ed country n_country
```

6. Now could you please find out the corresponding numbers for "Manufacturing" and "Telecommunications", "Greece", and "Luxembourg". Use command:

```
ed sector n_sector if sector == "Manufacturing"

ed sector n_sector if sector == "Telecommunications"

ed sector n_sector if country == "Greece"

ed sector n_sector if country == "Luxembourg"
```

7. How many years available in the dataset? Use command: sum year

8. Now our data is built across countries, sectors, and years (1995-2020). This is in particular an ideal panel dataset that is almost ready to be analysed. However, STATA still does not know how to treat this dataset as a "panel" that has a unique identifier to identify each "country-sector" pair.

   Suppose we are interested to analyse the impact of numbers of worked hours on the real value added and also real output through sectors and countries 1995-2020. As we have learned in the lectures, there are fixed variances that are time-invariant but variant across "countries", "sectors", and "years". We intend to use "Fixed-effects" estimator to get rid off these fixed-effects.

   So, the first thing is to tell STATA which variable is the "panel" identifier. Use command below:

```
sort n_country n_sector
egen panel_id = group(n_country n_sector)
```

   and then

```
sort panel year
ed n_sector n_country panel_id year
```

   now you should be able to see

| | n_sector | n_coun~y | panel_id | year |
|---|---|---|---|---|
| 24148 | Accommodation and food service activities | Finland | 929 | 2014 |
| 24149 | Accommodation and food service activities | Finland | 929 | 2015 |
| 24150 | Accommodation and food service activities | Finland | 929 | 2016 |
| 24151 | Accommodation and food service activities | Finland | 929 | 2017 |
| 24152 | Accommodation and food service activities | Finland | 929 | 2018 |
| 24153 | Accommodation and food service activities | Finland | 929 | 2019 |
| 24154 | Accommodation and food service activities | Finland | 929 | 2020 |
| 24155 | Activities of extraterritorial organisations and bodies | Finland | 930 | 1995 |
| 24156 | Activities of extraterritorial organisations and bodies | Finland | 930 | 1996 |
| 24157 | Activities of extraterritorial organisations and bodies | Finland | 930 | 1997 |
| 24158 | Activities of extraterritorial organisations and bodies | Finland | 930 | 1998 |

so, we now can see that for country ***Finland*** under ***Accommodation*** industry (***n_sector***) in year ***2019*** the panel identifier number (***panel_id***) is ***929*** and it is the same even in years 2014, 2015, 2016, etc. But once the industry (***n_sector***) changes (i.e., from accommodation to extraterritorial organisations and bodies) then the identifier number (***panel_id***) will change too from 929 to 930.

9. Now we can "xtset" our data. Use command:

**xtset panel_id year**

and you will see below:

*panel variable: panel_id (strongly balanced)*
*time variable: year, 1995 to 2020*
*delta: 1 unit*

10. Now Visiting the website: https://euklems-intanprod-llee.luiss.it/documentation/ and download the "variable list". Now consider the following model

$$y_{it} = \beta_0 + \beta_1 H_{it} + \varepsilon_{it}, \qquad \text{(EX57.1)}$$

where $y_{it}$ is the current prices of gross output (variable "***go_cp***" labelled as "***GO_CP***") for country-sector $i$ in year $t$, and $H_{it}$ is the total hours worked (by person engaged, variable "***H_EMP*** " labelled as "***h_emp***") for country-sector $i$ in year $t$. Using RE (random-effects) and FE (fixed-effects) estimators to estimate Eq. EX57.1. Use command:

**sum go_q go_cp h_emp h_empe**

**xtreg go_cp h_emp, fe robust**

**xtreg go_cp h_emp, re robust**

Discuss the results/outputs.

11. Which estimator is preferred? Why? Use command:

```
quietly xtreg go_cp h_emp, fe
estimate store FE1
quietly xtreg go_cp h_emp, re
estimate store RE1
hausman FE1 RE1
```

12. As mentioned in the lectures, in the level term real value vs nominal value matters a lot. To avoid this pitfall, we would like to estimate the following model instead:

$$lny_{it} = \beta_0 + \beta_1 lnH_{it} + \varepsilon_{it}, \qquad \text{(EX57.2)}$$

where $lny_{it}$ is natural log of the current prices of gross output for country-sector $i$ in year $t$, and $H_{it}$ is the natural log of total hours worked for country-sector $i$ in year $t$. Using FE (fixed-effects) estimators to estimate Eq. EX57.2. Use command:

```
sum go_cp
gen lngo_cp = ln(1+go_cp)
sum h_emp
gen lnh_emp = ln(1+h_emp)
```

13. While estimating Eq. (EX57.2) gives us the return of hours worked on output, the 2007 financial crisis globally may introduce disruptions that could be different across different countries and sectors. Using the knowledge learned in the lectures, a Difference-in-differences method may help.

If we group observations into two groups, one treated, one control; the treated group could be those who are sensitive to the global financial status, for instance, the financial sectors, manufacturing, etc. ***Agricultural***, ***mining and quarrying***, etc., under the other hand, are one of the industries that may be not dependent on the global economic environment. This gives us an idea that we could try to put sectors like Agricultural in a control group, and financial and manufacturing into the treated group.

Suppose we are interested in the following model:

$$lny_{it} = \beta_0 + \delta_0 d2_{it} + \beta_1 dT_{it} + \delta_1 d2_{it} \bullet dT_{it} + u_{it}, \qquad \text{(EX57.3)}$$

where $d2_{it}$ is a dummy variable for after treatment occurs, $dT_{it}$ is a dummy for whether treated, and $u_{it}$ is again the error term. How to estimate Eq.(EX57.3)?

Use command below:

```
ed sector n_sector if sector == "Agriculture, forestry and fishing"  //n_sector == 5

ed sector n_sector if sector == "Administrative and support service activities"
//n_sector == 4

ed sector n_sector if sector == "Arts, entertainment, recreation; other services and
service activities, etc"  //n_sector == 8

ed sector n_sector if sector == "Human health activities"  //n_sector == 17

ed sector n_sector if sector == "Human health and social work activities"
//n_sector == 18

ed sector n_sector if sector == "Other service activities"  //n_sector == 39

ed sector n_sector if sector == "Postal and courier activities"  //n_sector == 40

ed sector n_sector if sector == "Professional, scientific and technical activities;
administrative and support service activities"  //n_sector == 42

ed sector n_sector if sector == "Professional, scientific and technical activi"
//n_sector == 41

ed sector n_sector if sector == "Public administration, defence, education, human
health and social work activities"  //n_sector == 44

ed sector n_sector if sector == "Publishing, motion picture, video, television
programme production; sound recording, programming and broadcasting activities"
//n_sector == 45


gen dT = 1
replace dT = 0 if n_sector == 5
replace dT = 0 if n_sector == 4
replace dT = 0 if n_sector == 8
replace dT = 0 if n_sector == 17
replace dT = 0 if n_sector == 18
replace dT = 0 if n_sector == 39
replace dT = 0 if n_sector == 40
replace dT = 0 if n_sector == 42
replace dT = 0 if n_sector == 41
replace dT = 0 if n_sector == 44
replace dT = 0 if n_sector == 45

gen d2 = 1 if year > 2008
replace d2 = 0 if year <= 2007
```

14. Now, find out how many units are Treated, and how many units are Controlled; and how many Treated units are observed in the Before period, and how many Controlled units are observed in the After period?

Use the command below:

```
gen dTd2 = dT*d2
reg lngo_cp dT d2 dTd2, robust

tab dT d2 if e(sample)
su lngo_cp if d2 == 0 & e(sample)
su lngo_cp if d2 == 0 & dT == 1 & e(sample)
su lngo_cp if d2 == 0 & dT == 0
su lngo_cp if d2 == 1 & e(sample)
su lngo_cp if d2 == 1 & dT == 1 & e(sample)
su lngo_cp if d2 == 1 & dT == 0 & e(sample)
```

Please note that all the numbers should match the one provided by "tab dT d2 if e(sample)"

15. Now calculate $\bar{y}_{2,T} - \bar{y}_{1,T}$; the macro effects; the DiD estimator. Use the command below:

```
disp 10.47458 - 9.889637   //0.584943

disp 10.73719 - 10.06083   //0.67636

reg lngo_cp dT d2 dTd2, robust
```

```
Linear regression                               Number of obs     =     49,756
                                                F(3, 49752)       =     161.56
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0091
                                                Root MSE          =     3.2814

------------------------------------------------------------------------------
             |               Robust
     lngo_cp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          dT |  -.1711912   .0475372    -3.60   0.000    -.2643647   -.0780178
          d2 |   .6763629   .0573371    11.80   0.000     .5639815    .7887444
        dTd2 |  -.0914212   .0665891    -1.37   0.170    -.2219366    .0390942
       _cons |   10.06083   .0411144   244.70   0.000     9.980243    10.14141
------------------------------------------------------------------------------
```