

20269 – Economics of European Integration – 2024/25

Take Home

MAXIMUM ALLOWED SPACE = 25 PAGES INCLUDING TABLES

The Stata dataset `EEL_TH.2025.dta` can be downloaded from the course Blackboard. It contains data on more than 14,000 European firms operating in three countries (Spain, France, and Italy) and in two industries:

- **Manufacture of textiles** (coded as industry **13** according to NACE rev. 2).
- **Manufacture of motor vehicles, trailers, and semi-trailers** (coded as industry **29** according to NACE rev. 2).

The dataset covers the period **2000-2017**.

Variables Included in the Dataset

The dataset includes variables in levels, which are described below:

- **year** and **country**: Year and country of operation.
- **sector**: Industry classification according to NACE rev. 2 (codes **13** and **29**).
- **nuts2**: Region (**NUTS 2**) in which the firm operates.
- **id_n**: Unique identifier for each firm.
- **sizeclass**: Firm size class (based on the number of employees).
- **L**: Number of workers.
- **K**: Capital input (in thousands of Euros) and its deflated value (**real_K**).
- **M**: Raw materials input (in thousands of Euros) and its deflated value (**real_M**).
- **W**: Total labor costs (in thousands of Euros) paid by the firm.
- **sales**: Firm revenues (in thousands of Euros) and its deflated value (**real_sales**).
- **real_VA**: A proxy measuring the deflated value-added of the firm, calculated as:

$$\text{Value Added} = \text{Revenues} - \text{Materials}$$

Using this dataset, you are requested to solve the following problems:

Problem I

- Focus only on French firms.** Starting from balance-sheet data, provide some descriptive statistics (e.g. n. of firms, average capital, revenues, number of employees, value added) in 2007 comparing firms in sector 13 (textiles) and firms in sector 29 (motor vehicles, trailers and semi-trailers) for the region Nord - Pas de Calais. Please use the NUTS level 2 - 2013 definition. Comment briefly.
- Compare the descriptive statistics that you have analyzed in point a. for 2007 to the same figures in 2017. What changes? Comment and give an interpretation.

Problem II

- Consider now all the three countries.** Estimate for the two industries available in NACE Rev. 2 2-digit format the production function coefficients, by using standard OLS, the Wooldridge (WRDG) and the Levinsohn & Petrin (LP) procedure. How do you treat the fact that data come from different countries in different years in the productivity estimation?
- Present a Table (like the one below), where you compare the coefficients obtained in the estimation outputs, indicating their significance levels (*, ** or *** for 10, 5 and 1 per cent). Is there any bias of the labour coefficients? What is the reason for that?

	Nace-13	Nace-29
Lev-Pet		
ln(labor)		
ln(capital)		
WRDG		
ln(labor)		
ln(capital)		
OLS		
ln(labor)		
ln(capital)		
Bias in labour coefficient		
N. of observations		

Table 1: Comparison of Production Function Coefficients for NACE-13 and NACE-29

Problem III

- Would there be any difference in estimating the production function using revenues rather than added values in LP, WRDG or OLS? Why is it so? Discuss the issue theoretically, considering the assumptions behind the Cobb-Douglas production function.

Problem IV

- Comment on the presence of “extreme” values in both industries. Clear the TFP estimates from these extreme values (1st and 99th percentiles) and save a “cleaned sample”. From now on, focus on this sample. Plot the kdensity of the TFP distribution and the kdensity of the logarithmic transformation of TFP in each industry. What do you notice? Are there any differences if you rely on the LP or WRDG procedure? Comment.
- Plot the TFP distribution for each country. Are there any differences if you rely on the LP or WRDG procedure? Compare and comment.
- Focus now on the TFP distributions of industry 13 in France and Spain. Do you find changes in these two TFP distributions in 2006 vs 2015? Did you expect these results? Compare the results obtained with WRDG and LP procedure and comment.
- Look at changes in skewness in the same time window (again, focus on industry 13 only in these two countries). What happens? Relate this result to what you have found at point c.
- Do you find the shifts to be homogenous throughout the distribution? Once you have defined a specific parametrical distribution for the TFP, is there a way through which you can statistically measure the changes in the TFP distribution in each industry over time (2006 vs 2015)?

For the second part of this Take Home, you first need to recall how to compute the China shock at the regional level. To this purpose, you can refer to the formula by Colantone and Stanig (American Journal of Political Science, 2018):

$$ChinaShock_{crt} = \sum_k \frac{L_{rk}(\text{pre-sample})}{L_r(\text{pre-sample})} * \frac{\Delta IMPChina_{ckt}}{L_{ck}(\text{pre-sample})}$$

where c indexes countries, r regions, k industries in the manufacturing sector, and t years. You now need three more datasets (all downloadable from Blackboard).

The Stata dataset “Employment_Shares_Take_Home.dta” contains employment data for each NACE rev.1.1 industry in Italian, French and Spanish regions. As per the equation above, the employment variables refer to the earliest pre-sample year in which they could be observed. As such, they display the same values for each region and industry over time. The database is nevertheless structured as a panel (from 1988 to 2007) to facilitate the merge with the other two datasets presented below. You may use this dataset to compute weights for import deltas, i.e., the first term and the denominator of the second term of equation above.

The variables are included in the dataset in levels and are the following:

- **year** and **country** of each observation
- **nuts2**: the region code; **nuts2_name**: the region name
- **nace**: the NACE rev.1.1 industry
- **empl**: number of employees by region-industry pre-sample
- **tot_empl_nuts2**: total number of employees by region pre-sample
- **tot_empl_country_nace**: total number of employees by country-industry pre-sample

The Stata dataset “Imports_China_Take_Home.dta” contains data on imports from China to Italy, France and Spain for each NACE rev.1.1 industry in each year, from 1988 to 2007. You may use this dataset to compute import deltas, i.e., the numerator of the second term of the equation above. The variables are included in the dataset in levels and are the following:

- **year** and **country** of each observation
- **nace**: the NACE rev.1.1 industry
- **real_imports_china**: deflated imports from China

The Stata dataset “Imports_US_China_Take_Home.dta” contains data on imports from China to the US for each NACE rev.1.1 industry in each year, from 1989 to 2006. You may use this dataset to construct an instrumental variable, which is needed in the remaining part of the Take Home. The variables are included in the dataset in levels and are the following:

- **year** of each observation
- **nace**: the NACE rev.1.1 industry
- **real_USimports_china**: deflated imports from China to the US

Finally, the Stata dataset “EEL_TH_P6_2025.dta” contains data on European firms operating in 3 countries (Spain, France and Italy) over the period 2000-2017. You may use this dataset in Problem 6. The included variables are:

- **year** and **country** of each observation
- **nuts_code**: the region code
- **nace**: the NACE rev.1.1 industry
- **tfp**: mean TFP at the NUTS-2 and industry level
- **mean_uwage**: mean wage at the NUTS-2 and industry level
- **lnpop**: population in log
- **control_gdp**: GDP growth (%)
- **share_tert_educ**: share of population with tertiary education (%)

Problem V

- a. Merge the first three datasets together. Compute the China shock for each region, in each year for which it is possible, according to the equation above. Use a lag of 5 years to compute the import deltas (i.e., growth in imports between $t-6$ and $t-1$). Repeat the same procedure with US imports, i.e., substituting $\Delta IMPChina_{ckt}$ with $\Delta IMPChinaUSA_{kt}$, following the identification strategy by Colantone and Stanig (AJPS, 2018).
- b. Collapse the dataset by region to obtain the average 5-year China shock over the sample period. This will be the average of all available years' shocks (for reference, see Colantone and Stanig, American Political Science Review, 2018). You should now have a dataset with cross-sectional data.
- c. Produce a map visualizing the China shock for each region, i.e., with darker shades reflecting stronger shocks. Going back to the "Employment_Shares_Take_Home.dta", do the same with respect to the overall pre-sample share of employment in the manufacturing sector. Do you notice any similarities between the two maps? What were your expectations? Comment.
Note: A tutorial explaining how to create maps using Stata is available on youtube at the following link. Also, you can download the shapefile for the Nuts 2 map from the GISCO section of Eurostat website.

Problem VI

Use the dataset "EEL.TH.P6_2025.dta" to construct an average of TFP and wages during the post-crisis years (2014-2017). Create a lag of 3 years in the control variables (education, GDP and population). Now merge the data you have obtained with data on the China shock (region-specific average).

- a. Regress (simple OLS) the post-crisis average of TFP against the region-level China shock previously constructed, controlling for the 3-year lags of population, education and GDP. Comment on the estimated coefficient on the China shock, and discuss possible endogeneity issues.
- b. To deal with endogeneity issues, use the instrumental variable you have built before, based on changes in Chinese imports to the USA, and run again the regressions as in a). Do you see any changes in the coefficient?
- c. Now, regress (both OLS and IV) the post-crisis average of wage against the region-level China shock previously constructed, controlling for the 3-year lags of population, education and GDP. Comment on the estimated coefficient on the China shock. Lastly, what happens if you regress (both OLS and IV) the post-crisis average of wage against the region-level China shock previously constructed. Control for the average TFP during the post-crisis years, an interaction term between average of TFP and China shock and for the 3-year lags of population, education and GDP? Comment.

Problem VII

For the final part of this Take Home, you are asked to focus **only on Italian regions**.

First, please go to the European Social Survey official website and download the free-access Round 8 data. From this wave, keep only Italian respondents. Also, keep the following variables:

- Survey weights (post-stratification weight including design weight)
- Gender
- Age
- Highest level of education
- Region (NUTS 2)
- Party voted in last national election, which was held in the year 2013 (*prtvbit*);
- Support for the use of public money to subsidize renewable energies (e.g. wind, solar power) to reduce climate change (*sbsrnen*)

Second, please obtain the Manifesto Project dataset from the Manifesto Project website – available for free after registration. Then do the following:

- a. Merge the ESS dataset with data on the China shock (region-specific average), based on the region of residence of each respondent.
- b. Regress (simple OLS) the attitude score towards the allocation of public money to subsidize renewable energies on the region-level China shock previously constructed, controlling for gender, age, dummies for levels of education, and dummies for Nuts regions at the 1 digit level. Cluster the standard errors by Nuts region level 2. Be sure to use survey weights in the regression. Comment and discuss possible endogeneity issues.
- c. To correct for endogeneity issues, use the instrumental variable you have built before, based on changes in Chinese imports to the USA. Discuss the rationale for using this instrumental variable. What happens when you instrument the China shock in the previous regression? Comment both on first-stage and on second-stage results.
- d. Comment on the sign of the coefficient on the China shock obtained at point c. Provide intuitive theoretical arguments to rationalize this finding.
- e. Starting from the augmented ESS dataset used in the previous regressions, attach to the variable “party voted in last national election” the score for the “very general favorable references to underprivileged minority groups” available in the Manifesto Project for the election of 2013. You can search in the codebook the code for this score in the dataset. Please note that party names do not necessarily match exactly, and that ESS may have more parties than coded by the Manifesto Project. In the end, you should be able to match 9 parties.
- f. Regress (both OLS and IV, as above) the underprivileged minority groups score of the party voted against the region-level China shock, controlling for gender, age, dummies for levels of education, and dummies for Nuts regions at the 1 digit level. For this purpose, transform the score as follows:

$$\text{minority score}_{pc} = \log(.5 + z_{pc})$$

with z being the variable for the “very general favorable references to underprivileged minority groups”. Cluster the standard errors by Nuts region level 2. Be sure to use survey weights in the regressions. Comment.

- g. What can we learn from the results at point f on the relation between economic and cultural determinants of the globalization backlash?