Data Task 1: Universities

Please send a document with your results (in any format of your choice, e.g. Word, LaTeX, Slides, R Markdown, Jupyter Notebook...) as well as your code (preferably R, alternatively Stata or Python) until ***May 14, 2023*** *to leonie.bielefeld@econ.lmu.de and emilio.esguerra@econ.lmu.de. You can work in groups of up to three students.*

The goal of this data task is to replicate some key results from the following paper:

Cantoni, Davide, and Noam Yuchtman. "Medieval Universities, Legal Institutions, and the Commercial Revolution." *The Quarterly Journal of Economics*, Vol. 129, N. 2 (2014): pp. 823-887. DOI: https://doi.org/10.1093/qje/qju007

We will be discussing the main results from the paper together in Class 3 but you can already start working on the data task beforehand. You will find the paper as well as the necessary datasets on our Moodle course page.

1. Load the dataset `students_germany.csv` and reproduce Figure IV in the paper.
2. Load the dataset `analysis_germany.csv` that contains a city-year panel with 2,256 German cities between the years 1366 and 1406. Aggregate the dataset such that it contains a yearly time series with two variables. The first variable should contain the year and the second variable should contain the total number of newly established markets in Germany per 1000 cities in the given year.
3. Consider the following regression model:

$$Markets_t = \beta_0 + \beta_1 \cdot Year_t + \eta_t$$

   a) Run the above regression model using the time series dataset from (2), but only use the years **before 1386** in your estimation.
   b) Interpret the coefficient estimate for β_1.
   c) What is the predicted value $\widehat{Markets}_t$ in the year 1386?
   d) Run the above regression model using the time series dataset from (2), but only use the years **after 1386** in your estimation.
   e) Interpret the coefficient estimate for β_1.
   f) What is the predicted value $\widehat{Markets}_t$ in the year 1386?
   g) Calculate the difference between the obtained $\hat{\beta}_1$ from a) and d). Interpret this difference.
   h) Calculate the difference between the obtained $\widehat{Markets}_t$ from c) and f).

4. Reproduce the regression results from Column (1) of Table III in the paper by running the following regression model. Interpret the coefficient estimates for β_1, β_2, and β_3. (*Note: In order to receive the same results as in the paper, drop year "zero" (1386) from your sample.*)

$$Markets_t = \beta_0 + \beta_1 \cdot Year_t + b_2 \cdot Post_t + \beta_3 \cdot Year_t \cdot Post_t + \eta_t$$

5. Reproduce the regression results on the sample split from Columns (2) and (3) of Table III in the paper. To do so, you should start again with the main panel dataset `analysis_germany.csv`, calculate the median reduction in distance to a university and split the dataset into two parts along the median distance reduction. After that, you again need to aggregate the datasets, such that they contain one observation for entire Germany per year, respectively.

6. Reproduce the regression results for the placebo analyses in Table VII in the paper. You will need to use the datasets `analysis_england_wales.csv` and `analysis_italy.csv`. (*Hint: For calculating market establishment per 1000 cities, you need the number of cities, which you can find in the paper.*)

7. Produce a nicely formatted regression table with all of your regression results from exercises (4), (5), and (6). Your final regression table should hence include five columns:

   - Column (1): Main results from question (4) (Column (1) of Table III in the paper)
   - Columns (2) and (3): Sample split results from question (5) (Columns (2) and (3) of Table III in the paper)
   - Columns (4) and (5): Placebo analyses from question (6) (Columns (1) and (2) of Table VII in the paper)

Addendum: Tips for R

We recommend working with the open source software R to conduct the statistical analyses. To do so, you need to download and install R and R Studio. By installing R, you will install the programming language R on your computer. To work with R in a user-friendly way, you also need RStudio. RStudio is a so-called Integrated Development Environment or a clear interface for working with R. Both R and R Studio are available for free. You can download R here: https://cran.r-project.org. and R Studio here: https://posit.co.

You can find a lot of video tutorials on YouTube that are great for learning to get started with R, but also to find help with specific tasks. The documentation of packages and functions can be found on https://www.rdocumentation.org. The go-to resource for finding solutions to individual problems is Stackoverflow (https://stackoverflow.com/). Googling your coding problem or the received error message will often lead you to this widely-used forum.

You will need a number of functions from add-on packages for this data task (and most other empirical work):

Install the following packages and load them into your R environment:

- `tidyverse` (a powerful collection of packages for data cleaning, plotting...)
- `fixest` or `estimatr` (to run regressions with alternative standard errors)
- `texreg` or `stargazer` (to visualize and export regression output)

For example, install and load `tidyverse` by executing the following code:

```
install.packages("tidyverse")
library(tidyverse)
```

Here is a list of functions that might be useful for this data task:

read.csv()	Imports csv-files
ggplot()	Plots data
pivot_longer()	Brings data from wide to long format (opposite: pivot_wider())
group_by()	Aggregates data
summarise()	Calculates summary statistics of variables, for instance useful when aggregating data
ifelse()	Creates new variable based on values of other variable(s)
lm_robust()	OLS Regressions with robust standard errors
feols()	OLS Regressions, different SE possible and inclusion of fixed effects (if applicable)
summary()	Display regression output
texreg()	Export regression tables in tex format (use with feols)
stargazer()	Export regression tables (use with lm_robust)