Assignment 2

Ph.D. Applied Microeconometrics
KDI School Fall 2023

2024-10-01

Due date: Monday, October 14th at 6:59pm

For this assignment, you will be using the `village_economiccensus.rds` data in the assignment folder. I have done a bit of cleaning for you, including a variable `year` that is the year of the census. I have renamed all of the economic census variables to drop the `ec_` prefix. You can find the definition of the variables on the SHRUG website, from which I downloaded the data. The data dictionary for the economic census is here.

The dataset is in an R-specific format. You can load it into R using the `readRDS()` function.

Below is a list of tasks. I would like you to create a properly formatted PDF file, as if it were a paper. This means that raw code should not appear in the PDF file. When you estimate a regression, the choice of standard errors is up to you. However, I'd like you to justify your choice of standard errors in each case. For 3. and 4., please treat the data as a simple cross-section.

Tasks:

1. Create some new variables:
    - Proportion of total employment that is women
    - Proportion of total employment that is government employment
2. Create a table that shows the mean of these variables by year.
3. Create a figure of your choice. The choice of x and y variables are completely up to you. Describe the figure and interpret it.
4. Using OLS, estimate a regression (and present the output) that allows you to extract the exact same information as the information in the table you created in step 2. Please make sure to interpret the output.
5. Estimate a poisson regression that shows the relationship between the total number of female employees (on the left-hand side) and the following variables on the right-hand side: total employment, proportion of employment that is government employment, and year. Present the output of this regression. Please make sure to interpret the output.
6. The data also has an additional variable in it, called `shrid2`. This variable is a village identifier. It turns out that the data I've given you is *panel data*, with the exact same villages observed across four separate years. Re-estimate your regressions in 3. and 4. using this new information. How does this change any of the decisions you made in 3. and 4.? Please make sure to interpret the output and explain any changes you made.
7. Now, estimate a regression that shows the relationship between the proportion of jobs that are government jobs (on the left-hand side) and the following variables on the right-hand side:
    - Whether there is a coal plant within 50 kilometers
    - Whether there is a coal plant within 100 kilometers
    - Total employment (count)

I am interested in how much larger the effect within 50 km is than the effect within 100km is. In other words, if our regression equation is

$$y_i = \beta_0 + \beta_1 x_{within50} + \beta_2 x_{within100} + \beta_3 x_{emp} + \epsilon_i,$$

I am interested in the ratio $\frac{\beta_1}{\beta_2}$. Please bootstrap the standard errors of this ratio. Please present the output of this regression and the standard errors. I want two separate standard errors:

- One assuming there is no clustering on `shrid`
- One assuming there is clustering on `shrid`

As before, please submit the following files:

- Your R Markdown file
- Your knitted PDF file
- Any other scripts you used to complete the assignment