Problem Set 1: Panel Data

Joan Llull
Conghan Zheng

Microeconometrics
Fall 2024

Introduction

In this problem set, for each exercise you should **provide estimates, report standard errors, and briefly interpret the results**. Unless specified differently, you can use pre-programmed STATA or Matlab commands.

Due date September 27th 11 am, by email to conghan.zheng@uab.cat. You should submit in one zipped folder with:

1. a PDF document containing your analytical solution, if requested in the exercise, any output (figures, tables, etc.) obtained after running the code and your interpretations of the results, instructions on how to run your script(s) if necessary;
2. your code that could run without errors and reproduce the results reported in your explanatory document, specifying the package dependencies if necessary;
3. the name of the zip or rar file, which is `PS1+[your name]`.

1 Static Panel Data

Raw data in Figure 1 shows that the share of high-skilled immigrants over total employment increased substantially since the 1990s. From 2010 on, it grew more rapidly than the share of immigrants in low-skilled professions. Jaimovich and Siu (2017) reports that skilled immigrants are mostly concentrated in the so-called STEM occupations (Science, Technology, Engineering and Math). In this exercise, you will be estimating models adapted from Borjas (2003).

Data. The dataset `PS1_1.dta` was constructed using the *Survey of Income and Program Participation* (SIPP), a longitudinal survey designed to provide comprehensive data on income of American households and their participation in government income transfer programs. In this exercise, you will use the panel starting in 2008. `PS1_1.dta` includes individuals aged 25-60 employed in the private sector with a single job.

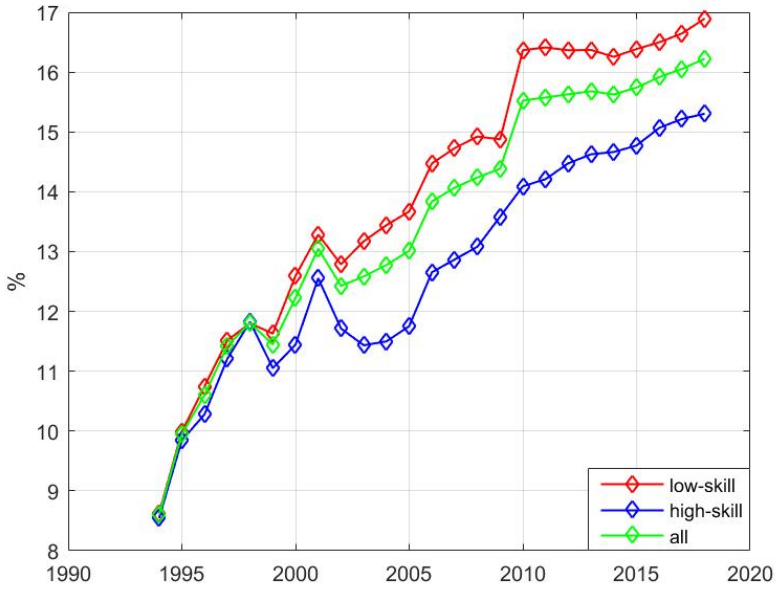

Figure 1: Share of immigrants in high- and low-skill employment, 1994-2018.
Source: Current Population Survey.

Model. This exercise consists of two parts: first, the estimation of the wage gap of skilled immigrants, and then the analysis of high-skilled immigrants' impact on US labor outcomes (wages and hours worked). Immigrants are defined as individuals born outside the US. High skilled workers are defined as those with a bachelor's degree or further education. The wage gap of skilled immigrants is given by β_1 in the following regression:

$$w_{it}^H = \beta_0 + \beta_1 immigr_i^H + X_{it}\gamma + \eta_i + \varepsilon_{it}, \tag{1}$$

where $immigr_i^H$ is a dummy variable that equals one if the individual is immigrant, X_{it} comprises worker characteristics such as age, gender and race, η_i is an unobserved worker fixed effect, and ε_{it} is the error term.

To estimate the impact of high-skilled immigration on outcome y_{jst} for job type j in state s at time t, we estimate the following regression:

$$y_{jst} = \theta p_{jst}^H + Z_{jst}\delta + \mu_j + \nu_s + \phi_t + (\mu_j \times \nu_s) + (\mu_j \times \phi_t) + (\nu_s \times \phi_t) + \xi_{jst}, \tag{2}$$

where μ_j, ν_s and ϕ_t capture job-type, state and time fixed effects, respectively. We distinguish four main types of jobs defined along two different dimensions: routine vs. non-routine and manual vs. cognitive.[1] The vector Z_{jst} contains characteristics such as the share of high-skilled workers, the share of females, etc. The parameter θ is meant to capture the impact of high-skilled immigration on outcome y_{jst}. The share of high-skilled immigrant workers p_{jst}^H is defined as:

$$p_{jst}^H = \frac{M_{jst}^H}{M_{jst} + N_{jst}}, \tag{3}$$

where M_{jst} (N_{jst}) is the number of immigrants (natives) in job j, state s at period t, and H denotes high-skilled workers.

[1]This classification follows Autor et al. (2003). Job types are *routine cognitive*, *routine manual*, *non-routine cognitive* and *non-routine manual*. In `PS1_1.dta` we assume that all jobs that are not cognitive are *manual*. Similarly, all jobs that are not routine are *non-routine*.

Exercise 1

1. Estimate the wage gap β_1 among high-skilled workers from Equation (1) using the *random effects* model. Remember that in order to identify β_1, the sample needs to contain only high-skilled individuals. Try also to estimate it using the within groups estimator. What happened? Why do you think that happened?

2. Compare your results from the previous section with those of OLS.

3. Test the *random vs. fixed effects* assumption. Explain what the test does. What do you think of your result?

4. Transform the data. Estimate Equation (2). Remember to use sample weights. Keep in mind that the dataset contains monthly observations, whereas Equation (2) requires yearly panel data, so you need to first aggregate your data. Interpret the fixed effects in your model. Also estimate (2) without the control variables in Z_{jst} using the within groups (or least squares dummy variables) estimator. Use both log wages and hours worked as y_{ijt}.[2]

2 Dynamic Panel Data

This exercise is based on Guner et al. (2018), which estimates the health gap between married and unmarried individuals. In a nutshell, their main findings stipulate that there is no effect of marriage on health for individuals younger than 40, whereas married above 40 years-old are healthier than the unmarried. Selection plays an important role here, as individuals with better health marry with higher probability. Using a subsample of their dataset, you will estimate some parts of their paper that relate to dynamic panel data.

Data. `PS1_2.dtax` contains a subsample of the data used in Guner et al. (2018). The data are taken from the *Panel Study of Income Dynamics* (PSID), a household survey that began in 1968 and constitutes a representative sample of 18,000 individuals from 5,000 households in the US. It provides information on income, employment, demographic characteristics and health outcomes. The dataset contains observations for household heads and spouses, both aged 20-64. The variable of interest (`healthy`) was constructed based on a categorical variable where individuals rate their health condition as either "excellent", "very good", "good", "fair" or "poor". It takes value one if either of the first three categories were reported and zero otherwise.

Model. The panel data structure of `PS1_2.dta` allows us to deal with unobserved heterogeneity and selection. The effect of marriage on health is assumed to vary over the life cycle. More precisely, let's first consider a simple fixed effects model:

$$h_{it} = \alpha(a_{it}) + \beta(a_{it})married_{it} + x'_{it}\gamma + \eta_i + \varepsilon_{it}, \tag{4}$$

where h_{it} is the health indicator for individual i at time t, a_{it} is the age interval, $\alpha(a_{it})$ is the age-dependent constant term (baseline health curve), $\beta(a_{it})$ is an additional component of the health curve associated with married individuals, and η_i is the individual fixed effect.

[2]Watch out: `hours_work` is coded as -8 when the working hours are varied.

The vector of additional controls x_{it} contains information on education, income, indicators for number of children, and year of birth. Using Equation (4), we can estimate the health gap between married and unmarried individuals over the life cycle. This specification takes into account selection coming from innate characteristics (individuals with better innate health may be more likely to marry). However, it does not take into account selection associated with health shocks (individuals that experience negative shocks may have lower probability of getting or staying married). To deal with the selection problem, we need to adopt a dynamic panel data model. Let's consider the following model

$$h_{it} = \phi h_{it-1} + \alpha(a_{it}) + \beta(a_{it})married_{it} + x'_{it}\gamma + \eta_i + \varepsilon_{it}, \tag{5}$$

that captures previous health shocks by controlling for health in the previous period h_{it-1}. All other variables remain the same as in Equation (4).

Exercise 2

1. Estimate Equation (4) using the within groups estimator. Remember to use survey weights. As control variables, you should include education, income, gender, indicators for the number of children, and year of birth dummies.

2. Estimate Equation (5) using the approaches proposed by Arellano and Bond (1991), and Arellano and Bover (1995). Try different specifications by changing the number of lags.

3. Graph the marriage health gap estimated in the two previous points and comment the results.

References

Autor, D. H., F. Levy, and R. J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 2003, *118* (4), 1279–1333.

Borjas, George J., "The Labor Demand Curve Is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market," *The Quarterly Journal of Economics*, 2003, *118* (4), 1335–1374.

Guner, Nezih, Yuliya Kulikova, and Joan Llull, "Marriage and health: Selection, protection, and assortative mating," *European Economic Review*, 2018, *104*, 138–166.

Jaimovich, Nir and Henry Siu, "High-Skilled Immigration, STEM Employment, and Non-Routine-Biased Technical Change," Working Paper, National Bureau of Economic Research 2017.