Midterm

Due by 4/8

**Generate a* `Python` *or* `R` *script for your answers. You will submit your script only.*

The midterm requires you to replicate the main results of Nollenberger, Rodríguez-Planas and Sevilla (2016). On Blackboard, you will find a pdf copy of the paper along with the data set, named `Final_sample.csv`. As such, you will produce Figure 1 on page 258 and Table 1 on page 260. Below you will find further instructions to replicate these results. Note that you will not be able to replicate their results exactly due to the fact that analysis of PISA data require some statistical methods beyond the scope of our class. Howbeit, your results should come close to theirs. Before start writing up your script, I urge you to read the paper carefully. It is a short paper and will not take too much of your time.

Consider the following research questions. What explains the observed lower average math scores of females relative to males? Are females less math oriented than males? If you were trying to answer the latter question, ideally you'd like to run a randomized controlled experiment where, hypothetically, gender can be randomly assigned to subjects who are the same in all other aspects. Then, one would simply compare average scores of the two groups. Since such an experiment is not feasible, we need to resort to observational data such as PISA. However, it is easy to find several confounding factors that explain math score and do systematically covary with gender. Such confounders need be controlled for in linear regression models. But problems do not end there. Consider institutions of countries or culture. Some countries have better education systems than others and some countries are culturally more gender-equal than others. As Nollenberger et al. (2016) state "*it is possible that greater gender equality leads to a reduction in the math gender gap, ... in countries where girls perform relatively better at math, women might also be more prepared, access better jobs, earn higher wages, and be more easily promoted and politically empowered, leading to greater gender equality.*" This is the so-called reverse causality problem. The authors' strategy to overcome this problem is to focus on second-generation immigrants (students) who have lived in a host country since birth, and are exposed to the same host-country institutions. These students will be exposed to the cultural beliefs of their parents' ancestry country. But note that the math test scores of these students are unlikely to affect culture or institutions of of their parents' ancestry country. Hence, the reverse causality problem is unlikely to occur.

Nollenberger et al. (2016) estimate different versions of the following specification:

$$\begin{aligned} pv1math_{ijkt} &= \alpha_1 female_i + \alpha_2(female_i \times ggi_j) + \boldsymbol{x}'_{ijkt}\boldsymbol{\beta}_1 + (female_i \times \boldsymbol{x}'_{ijkt})\boldsymbol{\beta}_2 \\ &\quad + \lambda_j + \lambda_k + \lambda_t + \delta(female_i \times \lambda_k) + \varepsilon_{ijkt} \end{aligned}$$

where $pv1math_{ijkt}$ denotes the (plausible) math test score of student i who lives in country k at time t, and is of ancestry j. $female_i$ is an indicator equal to one if student i is a girl and zero otherwise. ggi_j is the gender equality index from student i's country of ancestry of j. $\boldsymbol{x}_{ijkt}$ denotes a set of control variables which will vary depending on the specification considered. λ_j denotes the country of ancestry dummy, λ_k denotes the host country dummy, and λ_t denotes the PISA cohort dummy. They respectively control for time invariant country of ancestry characteristics, time invariant host country characteristics, and individual invariant cohort characteristics. Host country dummy is interacted with the female dummy to account for host country educational gender gaps. *The coefficient of interest is α_2, which captures the role of cultural on gender equality in explaining gender differences in the math test scores of second-generation immigrant girls relative to boys.*

Below you will find the description of the main variables used in the regressions in Table 1.

variable	description
pv1math	(plausible) math test score 1
ggi	gender gap index
female	indicator: 1 if female, 0 otherwise
age	age in years and month
diffgrade	indicator: 1 if the current individual's grade is different from the modal grade at the children age in the host country, 0 otherwise
misced	mother's highest level of education (categorical 0 to 6)
fisced	father's highest level of education (categorical 0 to 6)
momwork	indicator: 1 if mother works, 0 otherwise
dadwork	indicator: 1 if father works, 0 otherwise
lgdppc	log per capita GDP of the country
homepos	index of cultural possessions (positive values imply higher)
pcgirls	PISA index of the proportion of girls enrolled in each school
private	indicator: 1 if school is private, 0 otherwise
metropolis	indicator: 1 if school is a metropolitan area, 0 otherwise
background	parents' (both) country of birth
country	host country
stweight	sample weights to be used in regressions

(1) To replicate Figure 1 on page 258, you first need to calculate math gender gap values by country of ancestry (i.e., by *background*). To this end, you need to regress $pv1math$ on $female$ dummy by *background* country, and save the slope estimates for the female dummy. Then, you can generate a scatter plot where x-axis is the *ggi* of the ancestry country, and y-axis is the math gender gap estimates of the ancestry country from the regressions.

(2) To replicate Table 1 on page 260, you will estimate six different specifications. Pay attention to the set regressors in each specification. Note that in all specifications the dependent variables is $pv1math$. You need to include year fixed effects (λ_t), ancestry country fixed effects (λ_j), host country fixed effects (λ_k), and the interaction of female dummy with host country fixed effects ($female_i \times \lambda_k$) in all specifications except the third one, where there are no ancestry country fixed effects. Also, use sample weights *stweight* in all regressions,.

References

Nollenberger, N., Rodríguez-Planas, N. and Sevilla, A. (2016). The math gender gap: The role of culture, *American Economic Review* **106**(5): 257–61.