Exam - Applied Econometrics for Researchers, PhD

1 December 2021

Practical guidelines

The exam starts December 1, 2021, at 9.00 in K2.53.

Kindly bring your own laptop with Stata installed.

You will have access to the internet via eduroam and you may use all books, notes, etc. during the exam but you cannot help others or receive help from others during the exam.

Consider the assignments below. Respond to all questions contained in the assignments. More detailed answers, answers consistent with econometric reasoning, and clear accounts of actions are rewarded in the evaluation and grading. All assignments are solvable using the tools taught during the course.

For each problem, there is an indicative weight in the overall evaluation.

Files for the exam can be found in the folder "Exam" on the CANVAS page for this course.

You should use the provided .do-file AERexam.do. **Rename the file as AERfinal.do** and fill in own programming where appropriate.

At the end of the exam, you are kindly asked to hand in

- Your completed AERfinal.do file
- The .log file (AERexam.log) generated after running the complete program a final time
- A document with your final answers to each question, describing and interpreting the results as required in the assignments.

Hand in your answers at **13.00** at the latest. All hand-ins should be send to: hck.si@cbs.dk and to vr.si@cbs.dk

The exam is based on the study by Conti and Guzman (2021) on the US comparative advantage in entrepreneurship.[1] The authors investigate underlying sources of the US entrepreneurial ecosystem's advantage compared to other innovative economies by assessing the benefits Israeli startups derive from migrating to the US. Addressing positive sorting into migration, they show that "migrant startups" (i.e., startups that register their headquarters (HQ) in the US) raise larger funding amounts and also achieve a higher acquisition value. However, their patent output is not larger than that of "non-migrant startups" that remain in Israel.

Conti and Guzman conclude that the US entrepreneurial ecosystem's advantage vis-à-vis other innovative economies arises from several sources producing sizeable gains for startups, including investor availability and large consumer and acquisitions markets.

Data:

The dataset "AERexam_cross.dta" used in Question 1, 2 and 4 contains information on a cross-section of a total of 2,179 startups. Question 3 relies on a different (but related) panel data set of 1,895 startups observed over 4 years.

Please use the dataset "AERexam_cross.dta" to answer Questions 1, 2 and 4. The table below briefly describes the variables in the file:

Variable name	Variable description
us_hq	1 if the startup has established a US headquarter within 3 years after founding, 0 otherwise
total_raised	Total amount raised from VC (million US $)
ln_raised	= ln(**total_raised** +1)
n_founder	Number of founders
prof	1 if (one of the) founder(s) is a professor, 0 otherwise
n_ssfp	Number of prior successful startups of the founding team
cleantech	1 if industry is "Clean technologies," 0 otherwise
comms	1 if industry is "Communication technologies," 0 otherwise
it_software	1 if industry is "IT/Software," 0 otherwise
internet	1 if industry is "Internet," 0 otherwise
life_science	1 if industry is "Life sciences," 0 otherwise
medical_dev	1 if industry is "Medical devices," 0 otherwise
semicond	1 if industry is "Semiconductor," 0 otherwise
misc	1 if industry is "Hardware," 0 otherwise
CA_hq	1 if the startup moved its HQ to California, 0 otherwise
NY_hq	1 if the startup moved its HQ to New York, 0 otherwise
other_us_hq	1 if the HQ is set up elsewhere in the US, 0 otherwise
acquired	1 if the startup has been acquired, 0 otherwise
exit_amount	amount obtained when a startup is acquired (million US $)
incubator	1 if the startup has been part of an incubator, 0 otherwise
has_us_inventors	1 if the startup has at least one US inventor in the team, 0 otherwise

[1] Conti, A., Guzman, J., (2021), "What is the US Comparative Advantage in Entrepreneurship? Evidence from Israeli Migration to the United States", *Review of Economics and Statistics*, forthcoming. The replication data for the study can be downloaded from https://dataverse.harvard.edu/.

Question 1 (35%).

a) How many startups "migrated" to the US as reflected by the variable **us_hq**? What percentage of the total number of startups in the sample do “migrant startups” represent?

b) How many startups have raised venture capital (VC)? What is the mean value of **total_raised**, i.e. the amount of VC funding raised by these startups (in million US $)?

c) Draw a histogram of the total amount of VC funding raised by startups. Comment on the distribution.

d) Conti and Guzman (2021) consider the following transformation of the amount of venture capital raised after an initial round of funding: **ln_raised** = ln(**total_raised** +1). Why would they add 1 in this transformation? Draw the histogram of the variable **ln_raised**. Comment on the distribution and compare to the one you saw in Q1.c.

e) It is believed that startups that migrate to the US raise more VC funding due to greater investor availability. Do a t-test for significance of the difference between migrant and non-migrant startups based on the variable **ln_raised**. What is the estimated difference? Can you interpret this as the effect of moving a startup's HQ from Israel to the US, as indicated by the dummy variable **us_hq**, keeping all else equal? Why/why not? What is the direction of bias that you would expect?

f) Run a regression for **ln_raised** on **us_hq**, **n_founders**, **prof**, **n_ssfp** and industry dummies (note that there are 8 industry dummies: **cleantech, comms**, **it_software, internet, life_science**, **medical_dev**, **semicond**, **misc**). Interpret the coefficients that are statistically significant. Are there statistically significant differences between industries?

g) The funding premium obtained from moving a startup to the US could vary according to the state in which the HQ is set up. The data has dummy variables defined for different US destinations. Specifically, **CA_hq** equals 1 if the startup moved its HQ to California, 0 otherwise; **NY_hq** equals 1 if the startup moved its HQ to New York, 0 otherwise; and **other_us_hq** equals 1 if the HQ of a “migrant startup” is set up elsewhere in the US. Run a regression model in which you can test the hypothesis that startups that move to CA raise significantly more money than startups that move to NY. Do you find support for the expected difference in VC money raised between the two groups of firms?

h) Another belief from the literature is that the funding premium of moving your firm to the US increases with the founders’ prior success – is this true? Run a regression that allows the funding premium of migrant over non-migrant startups to vary with the number of prior successful startups (**n_ssfp**). Produce a plot to illustrate the relationship that you find.

Question 2 (25%).

In this question, you are asked to analyze the determinants of a startup's decision to migrate to the US. Conti and Guzman (2021) show evidence that there is "positive sorting into migration" and suggest that this observation could be an important input into the analysis of the potentially positive effect of the US entrepreneurship environment for the performance of startups.

a) Run a probit model for the probability to migrate to the US (i.e., **us_hq** = 1), based on the variables **n_founders**, **prof**, **n_ssfp** and **industry dummies**. Comment on the significance and sign of the coefficients.

b) How would you measure the average marginal effect of **n_ssfp** on the likelihood to migrate to the US based on the probit model obtained in Q2a? What would be the corresponding marginal effect estimate obtained from a linear model instead? Compare the two estimates.

c) Focus on the link between startup migration and VC funding as in Q1. Estimate the ATET of **us_hq** on **ln_raised** using propensity score matching. Include all the X variables that were found to be statistically significant in Q2a to construct matched samples.

d) Based on the model estimated in Q2c, check the balance condition for the propensity score matching estimator.

e) Based on the model estimated in Q2c, analyze the validity of the overlap (or common support) assumption.

f) When the analysis was reviewed for journal publication, Reviewer 2 was concerned with **us_hq** being endogenous, despite the efforts to match the samples. But the Reviewer did not specify why (which source of endogeneity is potentially at play here). Discuss briefly what could be their concerns, and if and how such concerns could be addressed by an instrumental variable strategy. What would be the required properties of an instrumental variable for this strategy to work in this case?

Question 3 (20%)

Reviewer 1 shared the same endogeneity concern but instead suggested to estimate a panel data model with firm fixed effects (FE). Their argument is that, since firms can be followed over time and funding can also change over time, a FE model would be an alternative way to consistently estimate the effect of the US entrepreneurial ecosystem on migrant startups.

a) Why would a firm FE model in this case be a potential solution for endogeneity of migrant status? Which assumptions need to be valid for this to be a credible strategy for identifying the causal effect of migration on startups' performance?

Following the Reviewer's suggestion, the results for this question are run on a dataset that covers the first 4 years (age 0, 1, 2, 3) of the lives of 1,895 startups. A total of 66 firms in this dataset are "early movers" in the sense that they have relocated to the US already by the founding year (age 0). The remaining 1,829 firms stay in Israel throughout their first 4 years of existence. These two groups are distinguished based on the dummy variable **earlymover** (equal to 1 for startups that migrated to the US by age 0; equal to 0 for non-migrant startups).

More precisely, the Reviewer suggested estimating a firm fixed effect panel data model that accounts for the migrant status of each startup, the current age of the startup, and the industry to which a startup belongs. The output below implements the Reviewer's suggestion and estimates the log of the cumulative amount of funding raised (**ln_cum_raised**) depending on the startup's **earlymover** status, **age**, and **industry**. (Note: Late movers are excluded for simplification).

b) Based on this output, what is the expected increase in funding between 1-year and 3-year old startups, if both stayed in Israel? And what is the expected increase in funding between 1-year and 3-year old startups, if both moved to the US? What do you conclude about the pace at which early movers and non-movers raise VC funding during their first years of activity?

c) Can you estimate the average funding premium (i.e., how much higher is the funding raised) between startups at age 0, comparing migrants to stayers, from the output below? Why/why not?

d) Explain why the coefficients of firms' industry affiliations are omitted from the output?

e) Reviewer 2 notes that different startups might have been founded in different (calendar) years. The reviewer asks you to control also for the year of foundation of the startup in the panel regression. Would you agree with the reviewer and go along with his/her suggestion?

f) Reviewer 3 notes that it would also be possible to estimate a random effects model. What would be the main advantages of doing this, and what assumptions are needed for this estimator to be consistent?

Output table for Question 3:

```
xtreg ln_cum_raised i.age##i.earlymover i.industry_code, fe r

note: 1.earlymover omitted because of collinearity.
note: 2.industry_code omitted because of collinearity.
note: 3.industry_code omitted because of collinearity.
note: 4.industry_code omitted because of collinearity.
note: 5.industry_code omitted because of collinearity.
note: 6.industry_code omitted because of collinearity.
note: 7.industry_code omitted because of collinearity.
note: 8.industry_code omitted because of collinearity.

Fixed-effects (within) regression               Number of obs     =      7,580
Group variable: firm_id                         Number of groups  =      1,895

R-squared:                                      Obs per group:
     Within  = 0.2471                                         min =          4
     Between = 0.1036                                         avg =        4.0
     Overall = 0.1177                                         max =          4

                                                F(6,1894)         =     134.10
corr(u_i, Xb) = 0.0861                          Prob > F          =     0.0000

                                (Std. err. adjusted for 1,895 clusters in firm_id)
------------------------------------------------------------------------------
                   |               Robust
   ln_cum_raised | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-----------------+------------------------------------------------------------------
             age |
              1  |   .2144036   .0116772    18.36   0.000     .1915021    .2373052
              2  |   .3432571   .0148633    23.09   0.000     .3141069    .3724074
              3  |   .4509673   .0171808    26.25   0.000      .417272    .4846625
                 |
    1.earlymover |          0  (omitted)
                 |
  age#earlymover |
            1 1  |   .4993587   .1116276     4.47   0.000     .2804327    .7182847
            2 1  |   .7628553   .1206916     6.32   0.000     .5261528    .9995578
            3 1  |   .9220344   .1311628     7.03   0.000     .6647958    1.179273
                 |
   industry_code |
 Communications  |          0  (omitted)
       Hardware  |          0  (omitted)
  IT / Software  |          0  (omitted)
       Internet  |          0  (omitted)
  Life Sciences  |          0  (omitted)
Medical Devices  |          0  (omitted)
  Semiconductor  |          0  (omitted)
                 |
           _cons |   .2826638   .0103126    27.41   0.000     .2624385     .302889
-----------------+------------------------------------------------------------------
         sigma_u |  .67426116
         sigma_e |  .38544398
             rho |  .75369985   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Question 4 (20%)

While VC funding is an important measure of startups' intermediate performance, there is also a considerable interest in their ultimate "exit" outcomes. These include being acquired by another firm or going public via an IPO (initial public offering).

For this part of the analysis, we consider again the full cross section of startups that you will find in the data file **AERexam_cross.dta**. The dummy variable **acquired** equals 1 for startups that have been acquired, and 0 otherwise.

a) What is the proportion of startups that have been acquired by other firms? How does this proportion vary with the migration status of the startup?

b) The variable **exit_amount** records the selling price, i.e. the amount obtained when a startup is acquired (in millions of US $). What is the average amount obtained by acquired startups?

c) Run an OLS regression of **exit_amount** on **us_hq**, **n_founders**, **prof**, **n_ssfp**, and industry dummies. What is the potential problem in interpreting the coefficient of **us_hq** in this regression as an estimate of the causal effect of startup migration on **exit_amount**?

d) We need an estimation strategy that can possibly correct for the fact that **exit_amount** is only observed for acquired firms (i.e., it is missing for firms that never get acquired). For this, Reviewer 2 suggests taking a closer look at two variables available in your dataset: **incubator** and **has_us_inventors**. The reviewer argues that these variables could be used as exclusion restrictions and recommends you to estimate a Heckman model to account for the potential selection bias in your sample of acquired firms. Follow the reviewer's suggestion and comment on the new estimated effect of **us_hq** on **exit_amount**. What do you conclude by comparing this with the estimate from Q4c?

e) What do you conclude regarding sample selection bias – is it a problem in this case? Which assumptions do you need to satisfy to trust the estimates of this model? Are these assumptions valid?