Firm Dynamics and Economic Growth
Problem Set I

Jose M. Quintero*

In this assignment you will replicate some of the key empirical findings of Akcigit and Kerr (2018) published version, from now on referred to as AK. There will be one important difference in the sample used for the analysis: while AK uses firm data from the Census, which is not publicly available, you will use a database of publicly traded firms, Compustat.

Your assignment is to replicate figures 1 and 2 from AK. Below are some guidelines for the assignment. NOTE: not all of the steps required to replicate the findings are listed below. You need to take some extra steps, which you will need to figure out yourself.

Optional: if you want to have more fun, you can look at the working paper version (NBER Working Paper #16443), which contains many additional empirical facts, and try to replicate any of them.

1 Download Data.

(a) Get an account on WRDS
(https://wrds-www.wharton.upenn.edu) in order to access Compustat.

(b) Access Compustat - Capital IQ - North America. Select the monthly-updated Fundamental Annual Database.

(c) Download the variables you need. Hints: select a date range that looks reasonable to you. Choose "gvkey" for company code. Select "Search the Entire Database". Select "EXCH" and "FIC". Select the variables "EMP", "SALE", "XRD", "SIC" in the query.

(d) You will also need patent data and you will need to match patents to their assignee firms in Compustat. The NBER provides a USPTO patent database, as well as a firm identifier already matched to Compustat. You can access patents, citations and Compustat merge identifier at https://sites.google.com/site/patentdataproject/.

2 Preliminary analysis.

(a) Sample Selection. Keep only firms incorporated in the US (hint: use "FIC") and firms that do business in U.S. dollars (hint: use "CAD").

---

*Full replication code can be found at https://github.com/jmquintero925/FD-EG-Ufuk

(b) Merge your Compustat sample with the patent database. To limit issues of truncation, limit your analysis to the year 1980-2000, i.e. only consider patents applied for ("appyear") between 1980-2000. What are the issues related to truncation when it comes to patents?

(S) Truncation creates is that patent application is a proxy for innovation flow. Thus trucating within a period will create noise around the innovation flow, favoring firms that applied to patents during the period of interest.

(c) Provide a table of summary statics (this will have to include some of the variables that you will compute in the next steps).

(S) The statistics for the full sample is presented in Table 1

Table 1: Summary Statistics

Variable	Obs.	Mean	Std. Dev.	Min	Median	Max
Employment	142,488	5.041	23.308	0.000	0.478	1,244
Sales	160,228	748.285	3,828.534	-80.817	54.635	206,083
R&D	73,230	23.926	188.644	-0.397	0.866	8,900
Innovative	184,206	0.299	0.458	0	0	1

Notes: Innovative is a dummy variable that takes the value of 1 whenever the firm has at least one patent during the period of interest.

(d) What is the share of firms which have at least one patent? We will refer to the firms with at least one patent as "innovative firms".

(S) As seen in Table 1 the share of innovative firms is close to 30% of the sample.

(e) Compare your sample of innovative firms to the population of Compustat firms. How do they differ? Think of a way to illustrate the differences (hint: for example, differences in employment size).

(S) The descriptive statistics controlling by the variable indicating if the firm has a patent during the period of interest is presented in Table 2 From Table 2 we see that there are significant differences between innovative firms and non-innovative firms. On average

(f) From now on, focus your analysis on the sample of innovative firms.

Table 2: Summary Statistics

	Not Innovative			Innovative				
	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.	$\mu_1 - \mu_0$	p-value
Employment	93,126	3.161	12.272	49,362	8.586	35.565	5.425***	0.00
Sales	107,863	453.263	1,950.590	52,365	1,355.978	6,038.598	902.715***	0.00
R&D	36,334	3.203	27.731	36,896	44.334	262.745	41.131***	0.00

Notes: *** denotes that the mean difference is statistically significant different than 0 at the 99% .

3   Firm Growth by Firm Size. (Figure 1).

(a) Build the main variable of employment growth used in the paper, i.e.

$EmpGr_{f,t} = (Emp_{f,+1t} - Emp_{f,t}) / Emp_{f,t}$. Apply the same winsorization discussed in the paper.

(b) Replicate Figure 1 and the corresponding regression. On the x-axis, you will use $Emp_{f,t}$. On the y-axis, you will use $EmpGr_{f,t}$. Hint: you can use the command binscatter.

(S) The replicated Figure is presented in Figure 1a. Just as in Akcigit and Kerr (2018) there is an negative relation between firm growth in terms of employees and their current employment.

Figure 1: Employment Growth by Firm Size

(a) Percentage Growth

(b) DHS Growth
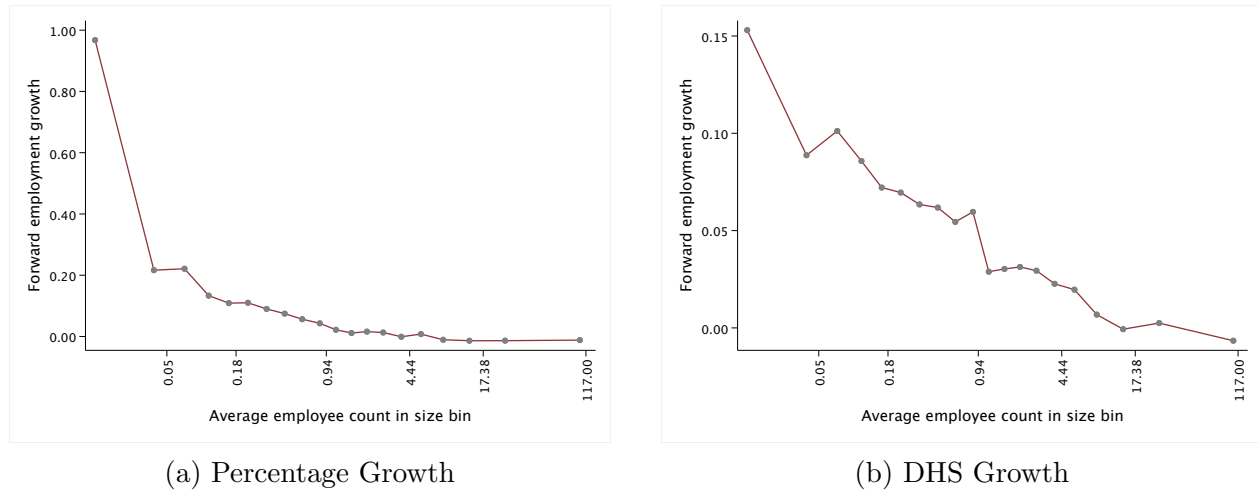
The regression corresponding to this exercise is presented in Table 3. Just like in the AK paper, there is a significant negative relation between employment and and firm growth. Specifically, an increase in one percent of the firm will decrease their growth by 0.08%.

(c) Replicate your graph for the alternative measure of firm growth rate cited in the paper, i.e.
$EmpGr_{f,t} = (Emp_{f,+1t} - Emp_{f,t}) / (0.5 (Emp_{f,+1t} + Emp_{f,t}))$.

Table 3: Firm Growth and Employment

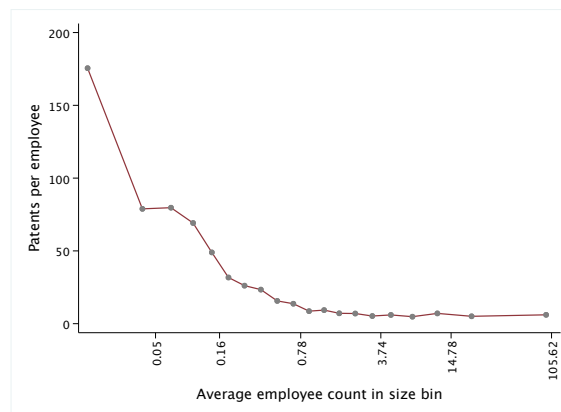| Variables	Employment Growth
log (Employment)	-0.0833***
	(0.0146)
Observations	46,295
R-squared	0.0169

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

(S) The results using the Davis, Haltinwanger and Schun measure for growth are presented in Figure 1b. Note that the relation is still the same but the scale is significantly different. The reason behind this change is that the DHS measure is absolutely bounded by 2 and thus the outlines will have a less meaningful impact.

4   Innovation Intensity by Firm Size. (Figure 2).

(a) Compute the number of patents per employment (i.e. total number of patents applied for in a given year by a firm, divided by employment). How lumpy is patenting in your dataset? Decide whether to apply the same five-year windows discussed in the paper and explain why.

(S) After inspection patenting is still lumpy. Conditional on the innovative sample, the median number of patents filed by a firm in a given year is 0. Moreover, on average, firms file for a new patent once every 1.5 years. Thus, I used a 5-year window as in the AK paper to smooth out this lumpiness.

(b) Replicate Figure 2 and the corresponding regression.

(S) The results are presented in Figure 2. As previously mentioned, I use windows of 5 years and include the 2000s in the last window.

Figure 2: Patent Intensity and Firm Size

As in the AK paper, small firms are more intense in patenting activity. The regression is presented in Table 4.

Table 4: Firm Growth and Employment

Variable	$\text{Patents}_{ft}/\text{Empt}_{ft}$
$\log\left(\text{Emp}_{ft}\right)$	-0.1187***
	(0.0090)
Observations	11,483
R-squared	0.1457

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Again, I find a robust negative correlation suggesting that small firms are more intense in their patenting activity.

5   R&D Intensity by Firm Size.

(a) In class, we discussed how the relationship between $R\&D$ intensity and firm size has evolved over time. We will replicate this finding here.

(b) Run a regression of $\log$ R&D expenditure over log sales in each year, with industry fixed effects :

$$\ln\left(R\&D_{i,j,t}\right) = \beta_0 + \beta_1 \ln\left(\text{ Sales }_{i,j,t-1}\right) + \delta_j + \epsilon_{i,j,t}$$

where i indicates a firm, j a sector and t a year. δ_j are sector fixed-effects. Your coefficient of interest is β_1. You will obtain a different coefficient β_1 for every year in your sample. Plot the coefficients you obtain over time, including the standard errors. Hint: You will have years on the x-axis and β_1 on the y-axis.

(S) The results for each regression are presented in Figure 3.

(c) How did the relationship between R&D intensity and firm size evolve over time? Explain.

(S) Over time, the relationship between sales and R&D has decrease although it has always been statistically significant. It is worth highlighting, during the second half of 1980s, the effect was relatively stable.

Figure 3: R&D and Sales
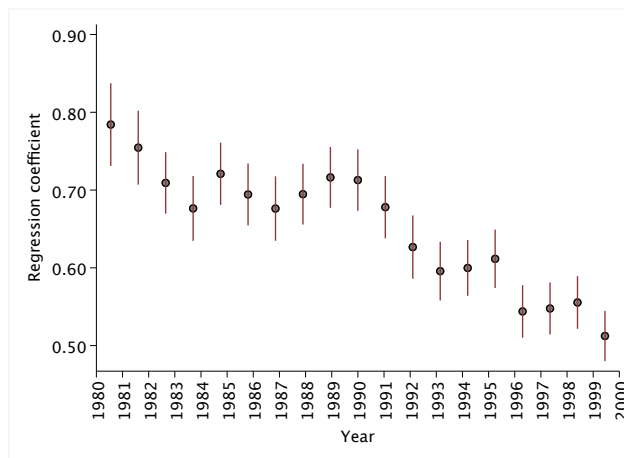
6 Discussion

(a) How do your results compare to the findings in the paper? Explain why.

(S) Overall, the results are consistent with the findings of Akcigit and Kerr (2018). Figure 1a and 2 replicate the results qualitatively. Similarly, the signs of regressions 3 and 4 are consistent with the results of the original paper. The difference in magnitudes is due to sample selection as firms showing up in Compustat are publicly traded companies.

(b) Do you believe that there are limitations in using Compustat versus Census data for the analysis? If so, what type of limitations?

(S) There is a significant bias in using Compustat versus the Census data. There is a considerable bias selection on publicly traded firms versus the entire universe of firms. Of course, firms that can innovate represent another margin of choice, which could be correlated with being a public company. Nevertheless, firms in Compustat are on average older, bigger and more capital/skill intensive. All these traits interact with incentives/ability to innovate and thus one has to account for these bias when restricting to Compustat.

(c) What do these findings imply? How are they useful to learn something about firm dynamics and innovation? How could you use them to inform a theoretical model?

(S) These results are particularly relevant as they marka departure from the stylized facts presented in Klette and Kortum (2004). For instance, in their model, patent intensity is constant across firm size. These fact is particularly relevant, as it highlights how as firms grow, their innovation incentives change.

Thus, as done by Akcigit and Kerr (2018), these moment can identify parameters of structural model and help design better industrial policy. To do so, these stylized facts create a mind teaser for the possible mechanism that might be acting and generating these results.