DRAFT - PLEASE DO NOT CITE

The Link between University R&D, Human Capital and Business Startups[1]

Nathan Goldschlag[2]
Ron Jarmin[2]
Julia Lane[3]
Nikolas Zolas[2]

Abstract

We expand the data infrastructure available to build evidence on public and private investments in science and R&D and utilize it to examine the links between startup performance and new measures of workforce human capital. We apply machine-learning techniques to a rich new source of longitudinally-linked data to characterize the research-experienced workforce of new businesses. Startups with a more research-experienced workforce are more likely to survive and grow.

[1] Disclaimer: Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. This research was supported by the National Center for Science and Engineering Statistics. NSF SciSIP Awards 1064220 and 1262447; NSF Education and Human Resources DGE Awards 1348691, 1547507, 1348701, 1535399, 1535370; NSF NCSES award 1423706; NIHP01AG039347; and the Ewing Marion Kaufman and Alfred P. Sloan Foundations. Data were generously provided by the Committee on Institutional Cooperation and its member institutions. We thank Cameron Conrad, Ahmad Emad, Christina Jones and Evgeny Klochikhin, for research support, Greg Carr, Marietta Harrison, David Mayo, Mark Sweet, Jeff Van Horn, and Stephanie Willis for help with data issues, and Jay Walsh, Roy Weiss, and Carol Whitacre for their continuing support. The research agenda draws on work with many coauthors, but particularly Bruce Weinberg and Jason Owen Smith.

[2] U.S. Census Bureau
[3] New York University

Introduction

There is a growing body of evidence documenting the decline in business dynamism over the past 30 years (Decker et al. 2014) characterized by a decline in both the formation and the success rate of new firms (Hathaway and Litan 2014). The reasons for the decline are not fully understood (Decker et al. 2016a), but the decline has important implications for resource allocation and productivity growth, especially in high-tech sectors (Decker et al. 2016b; Decker et al. 2017). Young entrepreneurial businesses are important for introducing and diffusing innovations in the economy and several authors have shown indirect linkages between formal investments in research and innovation and entrepreneurship and economic growth (Bania, Eberts, and Fogarty 1993; Hausman 2012; Lowe and Gonzalez-Brambila 2007).

Our point of departure for this study is the finding that the high-tech and information sectors exhibit different patterns of declining dynamism. Whereas most sectors exhibit a persistent secular decline, the high-tech and information sectors showed increases in dynamism during the tech boom of the 1990s followed by declining dynamism after 2000. Importantly, this pattern closely mimics that of productivity growth for the sector (Fernald 2014). Given the importance investments in science and R&D for driving innovation and growth in this sector, it is natural to ask whether changes in their scale, scope or impact can help explain trends in dynamism and productivity.

In this paper, we utilize a new and evolving data infrastructure that integrates detailed administrative data from research universities (Lane et al. 2015) with Census Bureau firm and worker data to begin to more directly investigate how university investments in science and research flow into the economy and impact entrepreneurship, productivity and growth. We construct new measures of human capital to investigate the contribution of worker experience with research to entrepreneurial success and dynamism.

We incorporate new worker-level measures of R&D human capital, including research training, of the workforce at both startups and young firms to directly examine the connection between an R&D trained workforce and new business success. We do not directly observe all these attributes for the entire universe of workers. Thus, we utilize machine-learning to scale our sample and generate estimates of the workforce with research experience. As such, an important contribution of the paper is demonstrating this new pilot approach to scaling and augmenting existing data collected at a local or regional level or for a subsample of firms and workers. Achieving these measurement objectives with survey data is not practical on both cost and respondent-burden dimensions

Our results suggest that a one-worker increase in the number of research-experienced employees in a startup firm's workforce increases its probability of survival to the next period by 1.9%, and increases the likelihood of it becoming a high-growth successful startup by 2.8%. Workers with experience in research increase the likelihood of startup success (defined as being 5-years old with ten or more employees) by 1.7%, over and above workers who have been employed by universities, high-tech or R&D-performing firms.

These results are consistent with the view that there is a relationship between workforce experience and business startup and survival. Further work using these data will be necessary to examine temporal dynamics. It will be particularly interesting to understand whether changes in the fluidity of this type of workforce, changing patterns of firm-to-firm job flows, or changes in the nature of research funding can be tied to the decline in business dynamism and changes in the distribution of employment growth rates.

Background

Decker et al. (2016b) review several studies that attempt to explain declining firm and labor market dynamism. Karahan, Pugsley, and Sahin (2015) find that changing demographics can explain declining startup rates, but Hyatt and Spletzer (2013) suggest that demographics play a limited role in explaining declining labor market dynamics. Changes in the industrial composition of the economy over time should have increased business dynamism as the share of activity accounted from by low volatility sectors like manufacturing was eclipsed by high volatility sectors such as services and retail (Decker et al. 2014). Finally, Goldschlag and Tabarrok (2014) find no evidence that increased federal regulations play any role in explaining trends in business dynamism, but Davis and Haltiwanger (2014) find labor market regulations have measurable impact of labor market fluidity.

The finding that trends in declining dynamism differ across sectors offers perhaps the best hope for pinning down causal factors (Decker et al. 2016c). The shift of retail from mom and pop stores to large national chains was the story over 1980s and 1990s. This is part of a long run trend that has been productivity enhancing. During the 1990s, businesses in the high-tech sector grew more dynamic with important implications of overall productivity growth. Since 2000, however, the high-tech sector exhibited large declines in measures of dynamism corresponding with weaker overall productivity growth.

Some have offered that slowing scientific discovery and innovation account for the slowdown in productivity growth (Bloom et al. 2016; Gordon 2016) or that ideas are not diffusing as efficiently as before and that the productivity gap between frontier firms and the rest of the economy is growing (Andrews, Criscuolo, and Gal 2015). From the viewpoint of models of firm dynamics (H. A. Hopenhayn 1992; H. Hopenhayn and Rogerson 1993) this should imply that the productivity shocks firms face have declined in magnitude and/or persistence. Decker et al. (2017) find that this is not the case but that firms' responsiveness to such shocks has decreased in a pattern that closely mimics patterns of dynamism and productivity growth. Relatedly, recent work by Gutiérrez and Philippon (2016) document a downward trend in business investment they ascribe mostly to decreased competition and risk averse institutional investors.

The finding that firms have become less responsive to productivity shocks suggests the presence of frictions. The pattern of increasing then decreasing dynamism in high-tech and the increasing dominance of frontier firms is suggestive that many firms in the economy are unable to identify and/or act on profitable opportunities arising from science and innovation. There is a literature that suggests that firm and economic growth can be significantly affected by workers specialized in R&D (Acemoglu et al. 2013; Jones 2002) that may be relevant especially for understanding the evolution of dynamism and productivity growth in the high-tech sector.

New linked data sets offer the potential to analyze the role of workers' and entrepreneurs' specialized training and experience. Particularly useful in this context is economy-wide linked employer-employee data, such as the LEHD data (Abowd, Haltiwanger, and Lane 2004). Such data have been used in the past to generate different measures of worker experience at different types of businesses (Golan, Lane, and McEntarfer 2007). Barth et al., for example, show that there are returns to experience at R&D performing firms (Barth, Davis, and Freeman 2016); Abowd et al. also use linked data to compute person specific measures of human capital (Abowd et al. 2005).

More direct measures of research human capital are now available, which include specific information on whether workers are trained in scientific research. The new longitudinally linked data on the research trained workforce - the UMETRICS data, (J. Lane et al. 2014) - have been used in other contexts and do suggest that research trained workers are more likely to work at firms with characteristics closely linked to productivity (Zolas et al. 2015).

Finally, there's an extensive related literature that links regional economic development clusters with the presence of active research universities (Glaeser, Kerr, and Ponzetto 2010; Hausman 2012; Kantor and Whalley 2013, 2014). The findings are consistent with the notion that an important source of knowledge transfer is the flows of research-experienced workers from one firm to another (Fleming, Charles King, and Juda 2007; Marx, Singh, and Fleming 2015).

Approach, Data and Measurement

Our framework posits that startup outcomes (Y) such as the survival and subsequent success of a startup *f* at time *t* is driven by capital (K) and technology (A), quantity and quality of labor measures (L) such as human capital, and external factors (X) such as macroeconomic conditions and industry factors. Functionally, we can think of outcomes being written as:

$$Y_{ft} = f\left(A_{ft}, K_{ft}, L_{ft}, X_{ft}\right) \quad (1)$$

For firm *f* at time *t*. We construct measures for each of these components using existing Census microdata on linked employee-employer data, longitudinal firm-level data, as well as existing surveys which indicate whether or not the firm is or was an R&D performing firm. We supplement this data with new data from UMETRICS, which identifies all workers who were paid on research grants for 14 universities that account for approximately 15% of federally funded research. Our primary focus is constructing components for the measure L_{ft}, which consists of the attributes of the startup workforce at time *t=0*.

Identifying Startups and Startup Outcomes

We create a Startup Firm History File (2005-2014) based on a panel database of age zero establishment attributes. The primary frame for the data is the Longitudinal Business Database (LBD), supplemented with additional information from the Census Bureau's Business Register, upon which the LBD is based. We utilize this file to identify startups by yearly cohort. Once the startups have been identified, we supplement the data with geocodes (state and county-level

FIPS, along with Census Tract information if available) and EINs taken from the Business Register. These variables are used to subsequently characterize the workforce associated with each startup gathered from LEHD (Longitudinal Employee-Household Dynamics) and W2 records. The full file contains data on employment, payroll, industry, geography, firm-type and birth/date of the firm.

Figure 1 below provides a graphical summary of the number of startups each year, including the share that fail in the subsequent years.

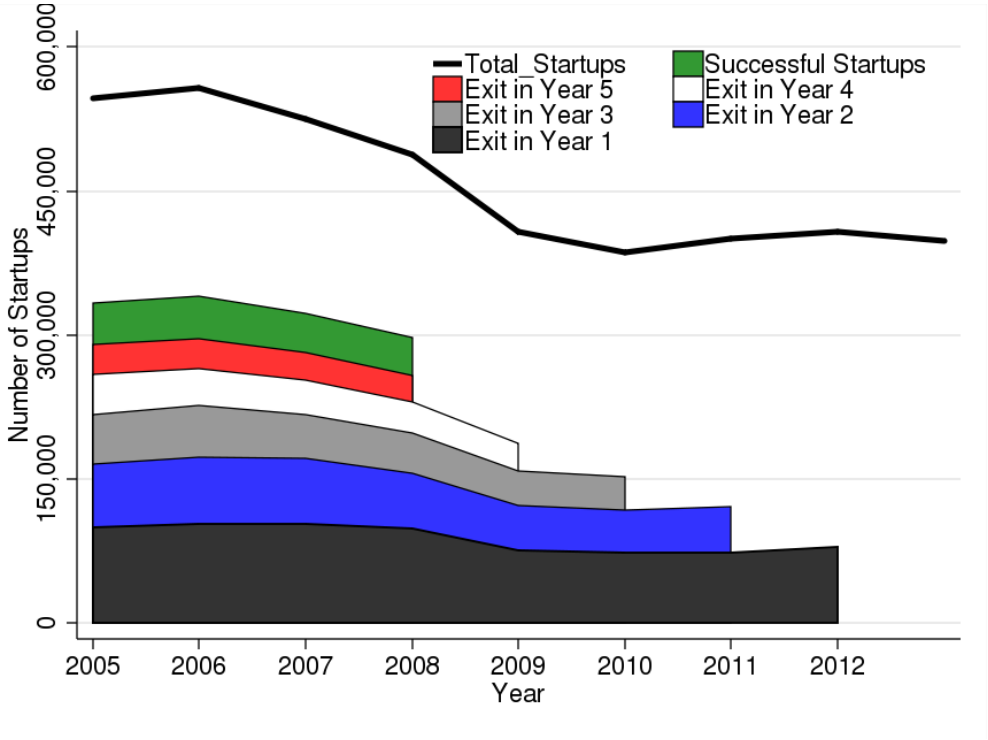


Figure 1: Number of Startups and their Death Rates first 5 years[4]

Figure 1 shows the counts of startups, as well as exits in each subsequent year, for the data sample. It also shows how many "successful" startups there are (defining success as surviving to year 5 and having more than 10 employees). Consistent with earlier findings, the number of startups declined by more than 25% from 2005 and 2013. More than 30% of startups fail before Year 2 and more than 50% of startups fail before Year 5. The rate of success for startups is approximately 8% each year, meaning that more than 90% of all startups in any year either die or fail to hire more than 10 employees within 5 years.

Characterizing the Startup Workforce

To characterize the workforce associated with each startup we create a Startup Worker History File (2005-2014) derived from worker level data on jobs. Universe data on jobs come from administrative records. Each paid job for each worker from 2005-2014 is reported at the Employer Identification Number (EIN) level via IRS form W2 and state-level Unemployment

[4] Source: Business Dynamic Statistics and Startup Firm History File.

Insurance wage records. The latter underlie the core LEHD infrastructure (Abowd, Haltiwanger, and Lane 2004) and are necessary to identify the establishment for the bulk of multi-unit firms (Abowd et al. 2009). The combined data includes more than 2.6 billion person-EIN-year observations (approximately 1.83 billion match across the W2 and LEHD/UI universes, 550 million are found only in the W2 records and 320 million are only found in LEHD). We then enhance this data with the LEHD Individual Characteristics File (ICF), which includes demographic data on persons including sex, age, race and place of birth.[5] We are able to link 48 million of the 2.6 billion person-EIN-year observations to startups in their birth year, giving us an average of nearly 4.5 million person-startup observations each year.[6]

We derive the human capital characteristics for each individual worker in the startup workforce at each time t from his or her work history in the previous three years. We create separate flags for whether the individual worked for (i) an R&D performing firm, (ii) a firm in a high-tech industry, (iii) a national research university and (iv) a national research university and paid on a research grant. The individual level data are then aggregated to create human capital composition measures for each startup for each year. The first three of these human capital measures are derived from a combination of different sources of internal Census Bureau data. The last is derived from new UMETRICS data combined with machine-learning methods as described below.

The R&D measure is created from adding firm-identifiers based on the Business Innovation and Research and Development Survey (BRDIS) and Survey of Industrial Research and Development (SIRD)[7]. A firm is classified as an R&D firm if it has positive R&D expenditures during the year the employee was affiliated with the firm. The high-tech industries classification is derived from work by Hecker (Hecker 2005; Goldschlag and Miranda 2016), which is based on the relative concentration of STEM workers. The university measure is derived from data from IPEDS and the Carnegie Institute which provide a frame of universities in the United States. We also merge in national university research outlays collected by National Center for Science and Engineering Statistics at the National Science Foundation and keep the top 130 universities that comprise of 90% of total federally funded R&D research.

The identification of individuals working on research grants can be derived from UMETRICS data (Lane et al. 2015), which includes 14 universities accounting for 15% of federally funded research. The UMETRICS data are universe data from the personnel and financial records of universities. Although four files are provided by the university, the key file of interest in this project is the employee file. Briefly, for each funded research project, both federal and nonfederal, the file contains all payroll charges for all pay periods (identified by period start date and period end date) with links to both the federal award id (unique award number) and the internal university identification number (recipient account number). In addition to first name and last name, and date of birth, the data include the employee's internal de-identified employee

[5] A detailed discussion on the matching process and match rates is provided in the appendix.
[6] This figure differs from the reported Business Dynamics Statistics (BDS), which calculate employment at startups at a specific point in time (March 12). Our figures are higher, reflecting employee-employer transitions (i.e. workers who work briefly for a startup and then move to a different job). The 48 million observations represent 37.8 million unique individuals.
[7] We use the SIRD to identify R&D firms between 2005-2007 and BRDIS for 2008-2014

number, and the job title (which we mapped into broad occupational categories). Each university provided data as far back as they had reliable records (see Appendix for more details). We extend the measure to all universities and back to 2005 using machine-learning approaches; that is discussed in the next section.

Machine-learning and Identifying Workers funded from research grants

The current UMETRICS frame consists of 14 large research universities, with several concentrated in the Midwest. Although some have provided data from the early 2000s, the bulk provide data for the latter years of our sample. The current UMETRICS frame consists of 140,000 research trained individuals that can be linked to Census data and used to create a training dataset for machine-learning purposes.

The training dataset consists of the employment and earnings records of all 14 UMETRICS universities in the period in which they provide data. By combining the UMETRICS and W2 data, we can identify all 140,000 who were employed on research grants in those time periods as well as 1.4 million who are not. The out-of-sample set includes 6.8 million individuals paid by the top 130 research universities in our time frame. Importantly, the out-of-sample set includes years for some UMETRICS institutions outside of those provided by the universities.

The link to Census data enables us to create a rich set of attributes that can be used to train the machine-learning models. We are able to capture each employee's earnings history before, during and after the employee's time at the university. In addition, we capture other attributes such as the dominant employer characteristics (includes size, payroll, average earnings, industry, location and other-job earnings), in-state and out-of-state earnings, industry earnings, geographic variation (across all 50-states), university characteristics (collected from IPEDS, Carnegie Institute, NSF and NIH, which include average SAT scores, enrollment levels, public/private indicators), along with yearly variations and before/after/during (for the period t-2 until t+2 for the individual entering and exiting the university) across all variables. All of this is supplemented with demographic data collected from the Individual Characteristics File (ICF). In total, we have over 1,500 person-EIN level features to train the machine-learning algorithms.

The success of our machine-learning methods hinges on the extent to which there are measurable differences between research trained and non-research trained individuals. Table 1 below highlights some key differences between employees working on research grants and those not.

Table 1: Comparison of demographic and earnings characteristics

	Research Trained	Not Research trained
Proportion Female	50.5	54.1
Proportion White	73.2	77.2
Proportion Hispanic	4.3	4.9
Proportion Black	5.7	9.3
Proportion Asian	14.1	6.2
Proportion Foreign-Born	21.8	11.4
Year of Birth	1977.7	1975.6
Proportion in Professional/Scientific Services	18.4	14.3
Professional/Scientific Earnings, t+1	42,500	33,700

Source: W2 and UMETRICS data.
Note: Each of these are significantly different at $p<0.001$.

Research trained individuals tend to be disproportionately male, Asian, foreign-born and younger relative to non-research trained employees (employees at the same institution but not affiliated with research grants). Research trained individuals are also more likely to be employed in Professional and Scientific services subsequent to leaving the university and have an earnings premium that is 30% higher in Professional and Scientific services in the year immediately following their exit from the university.

The quality of our classification methods also depends on the extent to which our UMETRICS universities are broadly representative of the 130 out-of-sample research universities. Table 2 compares the national university sample with the UMETRICS sample. The majority of universities included in the sample are large, public universities with medical schools attached to them. The UMETRICS sample is slightly larger on average and expends more on R&D. There are approximately 6.8 million Out-of-Sample individuals employed at these universities between 2005 and 2014.

Table 2: Comparison of university characteristics

	130 Universities	UMETRICS Sample[8]
Mean R&D Expenditure ($000), 2014	424,600	661,700
Mean Non-R&D Expenditures ($000), 2014	20,400	35,800
Mean # of NIH Awards, 2014	270	440
Mean Annual Enrollment	30,800	43,400
Mean Amount of NIH Awards ($000), 2014	112,500	180,900
Mean Undergraduate Enrollment, 2014	19,800	27,700
Mean Bachelor Degrees Awarded, 2014	4,700	6,900
Mean Graduate Enrollment, 2014	7,900	11,800
Mean Master Degrees Awarded, 2014	1,900	3,100
Mean Doctoral Degrees Awarded, 2014	700	1,100
Mean Total Degrees Awarded, 2014	7,300	11,100
Mean Faculty Number, 2014	1,400	2,200
% Private	28.5	30.8
% Land Grant	40	61.5
% with Medical School	69.2	84.6
Mean SAT Combined, 2014	1,140	1,190

Source: IPEDS and the Carnegie Institute.

The objective of our machine-learning approach is to classify individuals in the out-of-sample set as to whether or not they participated in (were paid by) grant funded research. Our methodology proceeds as follows. First, we execute several feature selection models. Second, we estimate a series of supervised learning classification models with different parameterizations. Third, we perform a number of cross validation exercises to assess the sensitivity and robustness of the in-sample predictions. Finally, we use our preferred specification to predict which of the 6.8 million out-of-sample individuals participated in grant-funded research.

We perform a series of feature selection exercises to reduce the number of attributes considered by each learning model. Feature selection can provide a number of benefits including avoiding over-fitting, reducing computational burden, and improving prediction quality by filtering low value added features and/or selecting a subset of the most valuable featured based on prediction quality (Guyon and Elisseeff 2003). We explore several univariate feature selection methodologies including k-best chi squared and univariate k-best by decision tree precision. We also use mean decreased impurity in a multivariate random forest model (Kohavi and John 1997). Finally, we develop some hand-curated feature sets based on iterative implementation and testing. Each of the resulting feature subsets are used to train the classification models.

For each of the k-best methods we select the top 50 features.[9] The k-best chi squared method estimates the chi-square test statistic between each feature and class (research training status) and selects the top *k* features based on those estimates. This method measures the dependence between each feature and class removing those that are most likely to be independent of research training status and therefore less useful for classification. The k-best decision tree method

[8]Ohio State University, Penn State, Purdue, Michigan State, New York University and the Universities of Arizona, Illinois (Champaign-Urbana), Iowa, Michigan, Missouri, Wisconsin.

[9] In the future, we plan to experiment with the 100 and 200 best by each method.

estimates a decision tree classifier for each feature and class individually and evaluates the quality of in-sample predictions based on that single feature. Intuitively, features that have less predictive value will produce lower quality predictions when used in a univariate classification model. Features are ranked according to the mean stratified three-fold cross-validated precision score from fitting the decision tree classification model for each feature-class combination. Precision, discussed in more detail below, captures the probability that a randomly selected positive predicted research training status is true. For our purposes, precision is the most relevant measure since we are most interested in measuring economic outcomes associated with positively classified individuals in the out-of-sample set.

Multivariate feature selection methods improve upon univariate methods by incorporating the complex interactions that can occur between features in supervised learning classification models. We calculate the k-best features by mean decreased impurity (Gini importance) in a random forest classifier (Breiman et al. 1984). The Gini importance measure is derived from the Gini index used to split the data at each node, which captures the level of impurity/inequality among samples assigned to a node based on the split from its parent node (Zhang and Ma 2012).

We estimate several classification models including logistic regression, decision tree, and random forest. The first classification model we estimate is a logistic regression classifier, a classic supervised learning method for binary classifications problems (Fan et al. 2008; James et al. 2013). This model serves as a baseline from which we compare the performance of the tree-based methods. The second classification model we estimate is the decision tree model (Breiman et al. 1984). Finally, we estimate a series of random forest models with different parameterizations (Breiman 2001). We explore a series of evaluation metrics for in-sample predictions resulting from different parameterizations of the random forest classifier.

To evaluate our predictions, we calculate several quality measures including accuracy, precision, and recall. We also use the share of false positives and false negatives to guide model selection and parameter tuning. Accuracy captures how often the model is correct with respect to both positive and negative classifications. This measure will tend to be less useful for our purposes since we are most concerned about correctly identifying positives (those that participated in grant funded research). Accuracy is defined in the following way.

$$ACCURACY = \frac{tp + tn}{tp + tn + fp + fn} \tag{2}$$

Where *tp*, *tn*, *fp*, and *fn* are true positives, true negatives, false positives, and false negatives respectively. Precision can be thought of as the probability that a randomly selected person predicted to have participated in grant funded research actually did. Recall, on the other hand, captures the probability that a randomly selected grant funded researcher was correctly classified. Since we are primarily interested in the quality of positive classifications, in the discussion that follows precision will be our primary measure of quality.

$$PRECISION = \frac{tp}{tp + fp}$$
$$RECALL = \frac{tp}{tp + fn} \quad (3)$$

The accuracy, precision, and recall measures estimated by training and predicting using the entire training set will suffer from over-fitting. To avoid this issue, and obtain a more accurate measure of model quality, we perform several cross validation exercises. First, we execute a stratified K-fold cross validation strategy. Second, we perform "Leave-One-Out" cross validation at the university level.

Using stratified K-fold cross validation, we segment the data into 10 folds stratified in such a way that each sample contains approximately the same relative frequency of observations within each class (research trained (1s) and non-research trained (0s)). We then cycle through each fold, training the classification algorithm using the K-1 samples and test on the Kth. For the "Leave-One-Out' cross validation we iterate over the UMETRICS universities leaving one out, training the model using the remaining universities and predict on the excluded university. This allows us to simulate the addition of a new university to the UMETRICS data.

Table 3 shows the in-sample evaluation metrics for the logistic regression and decision tree classification models using several feature selection sets.

Table 3: Logistic Regression and Decision Tree Classification Results

	Logistic Regression			Decision Tree		
Feature Set	Chi-Squared	Decision Tree	Impurity	Chi-Squared	Decision Tree	Impurity
In-Sample Accuracy	88.405	88.431	88.422	99.984	97.847	99.996
In-Sample Precision	32.090	30.000	33.566	99.991	99.254	99.995
In-Sample Recall	0.274	0.064	0.170	99.872	81.987	99.970
Mean 10-Fold Precision	30.034	36.656	36.852	28.158	27.386	31.542

Source: UMETRICS, W2, LEHD, LBD, ICF and BR.

The results in Table 3 show that while accuracy is relatively high with the logistic regression classifier, it generally fails to predict research trained individuals with precision of roughly 31 across the different feature sets. Moreover, the recall for the logistic regression model is very poor. The decision tree results for all three feature sets, on the other hand, appear very promising with nearly perfect accuracy and precision. However, in the stratified 10-fold validation we see that while the logistic regression model retains its precision scores of roughly 30 in the cross validation, the decision tree model performs significantly worse in cross validation. This suggests that the decision tree tends to over-fit the training sample. Table 4 shows the results from the random forest classifier across the feature sets.

Table 4: Random Forest Classification Results

	Random Forest			
Feature Set	Chi-Squared	Decision Tree	Impurity	Hand-Curated
Estimators	50	50	50	50
Maximum Depth	50	50	50	50
In-Sample Accuracy	99.703	97.277	99.950	99.924
In-Sample Precision	99.990	98.971	99.991	99.981
In-Sample Recall	97.443	77.248	99.580	99.365
Mean 10-Fold Precision	99.849	70.816	99.806	62.222
Mean University-Fold Precision	86.873	86.661	86.710	

Source: UMETRICS, W2, LEHD, LBD, ICF and BR.
Note: Fewer than 50 estimators and lower maximum depth results in significant loss of precision and accuracy while additional estimators and depth yield little additional quality improvements and entail significant additional computational resources. The hand-curated set includes demographic variables and demeaned earnings for individuals during their time at the university.

The results in Table 4 suggest that the random forest classifier, by aggregating many different decision trees, avoids some of the over-fitting issues in the decision tree results. The accuracy, precision, and recall across the different features sets are high, with exception of the recall score for the univariate decision tree feature set, which drops from over 97 to about 77. This pattern is also evident in Table 3, where we see the univariate decision tree produces lower recall scores for both the logistic regression and decision tree classifiers. We also show in Table 4 the results using a hand-curated set of features, which includes demographic variables and demeaned earnings. We create this hand curated set by iteratively experimenting with different combinations of features to balance the quality of in-sample predictions with the number of out-of-sample positive predictions.

Applying our preferred classification model, the random forest estimator with the hand-curated feature set, to the out-of-sample set identifiers an additional 188,000 individuals who are likely to be research trained. Table 5 below compares the out-of-sample results with the in-sample and individuals not likely to be research-trained.

Table 5: Comparison of Economic and Demographic Characteristics

	Research Trained		Not research trained
	In Sample	Out of Sample	
Proportion Female	50.5	47.8	54.1
Proportion White	73.2	67.6	77.2
Proportion Hispanic	4.3	7.1	4.9
Proportion Black	5.7	5.6	9.3
Proportion Asian	14.1	16.2	6.2
Proportion Foreign-Born	21.8	25.6	11.4
Year of Birth	1977.7	1976.2	1975.6
Proportion in Professional/Scientific Services	18.4	18.4	14.3
Professional/Scientific Earnings, t+1	42,500	41,250	33,700

Note: Each of the differences listed in this table are statistically significantly different at $p<0.001$.

The out-of-sample prediction of research training compares favorably with the known in-sample group of research trained individuals. There are a couple of notable differences however. The out-of-sample is significantly more likely to be male, Hispanic and foreign-born than the in-sample.

Our combined data consists of a national sample of Startups and their outcomes between the years 2005 and 2014, as well as a national sample of all workers affiliated with these startups, along with 4- main designations of human capital attributes assigned to each worker. The next section explores some basic summary tables and findings for these different types of workers and their potential impact on startups.

Basic Facts

This section establishes some basic facts on the human capital composition of the startups by year, as well as startup outcomes. We begin by outlining the data construction for the human capital element of the Startup Firm History File.

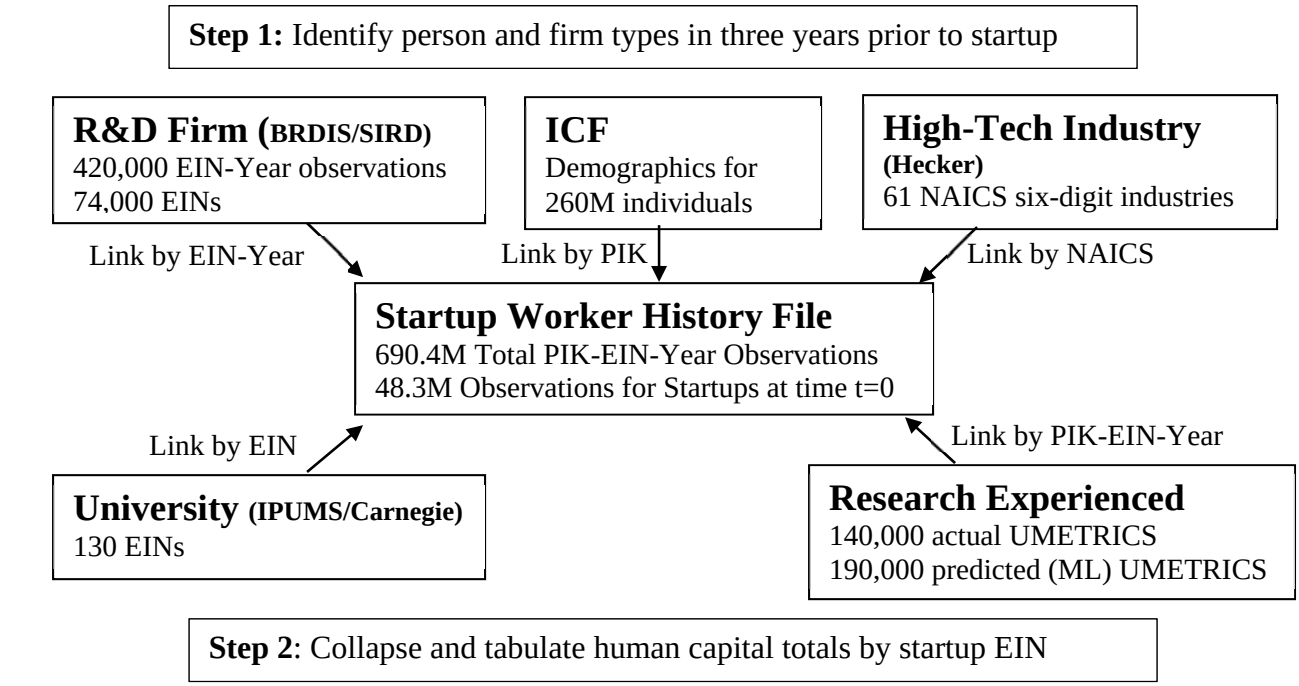
We start with the Startup Worker History File which consists of an individual protected identification key (PIK), their affiliated EIN, Earnings and Year for all individuals affiliated with startups at time t=0. We then affix human capital identifiers to their work history using the BRDIS/SIRD (which identifies R&D performing firms), High-Tech firm classification (from Hecker (2005) and Goldschlag and Miranda (2016)), University identifiers (from the 130 top research institutions whose EINs are collected by IPEDS and the Carnegie Institute) and our "research-experienced" measure compiled in the previous section. We then classify individuals into these four categories based on whether or not they worked at one of these institution types in the 3-years prior to them working at the startup at time t=0. We also merge in demographic information from the Individual Characteristics File (ICF) by PIK. We then sum up the totals of each of the human capital categories by startup EIN at time t=0 and merge these onto the Startup Firm History File.

The Startup Firm History File contains all non-farm private business that sprung into existence beginning in 2005 and charts their employment history through 2014 (or until firm death). The

file contains industry codes, employment, revenue and payroll. Using the industry classification, we can identify high-tech firms and "industrial" firms.[10] We then classify a number of different outcome measures, such as survival, success and high-growth success (classified as either 1/0 for each startup). A depiction of the file construction is below in Figure 2.

[10] We classify "industrial" firms as startups with potentially greater value-added to the economy than basic service industries. These include manufacturing startups (starting two-digit NAICS 31-33), Information (starting two-digit NAICS 51), Finance and Insurance (starting two-digit NAICS 52), Professional, Scientific and Technical Services (starting two-digit NAICS 54) and Health Care and Social Assistance (starting two-digit NAICS 62).

Figure 2: Data Construction for Human Capital Measures of Startups

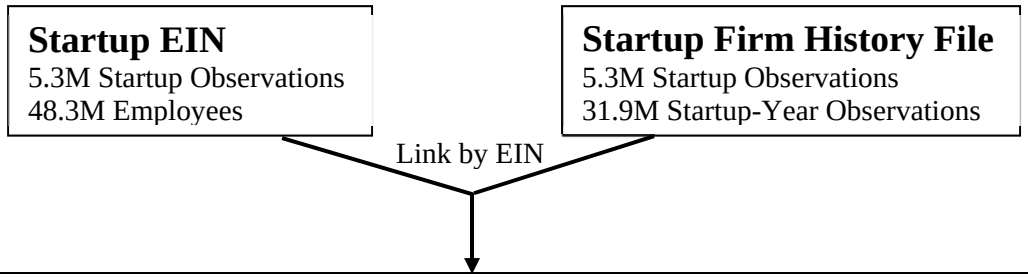


Step 2: Collapse and tabulate human capital totals by startup EIN

Startup EIN
5.3M Startup Observations
48.3M Employees

	R&D Firm	High-Tech	University	Research Experience	Female	Foreign Born
Startup Count	1.6M	900,000	400,000	35,000	2.8M	1.7M
Employee Count	8.1M	3.0M	950,000	50,000	22.5M	8.4M

Step 3: Merge with Startup Firm History File and classify startup types and outcomes at time t=0



	Startup Count	Survive to Next Period	Successful Startups	High-Growth
All	5.3M	3.4M	170,000	29,000
At least 1 R&D Firm Employee	1.6M	1.36M	120,000	20,000
At least 1 High-Tech Employee	900,000	780,000	75,000	12,000
At least 1 University Employee	400,000	360,000	41,000	6,300
At least 1 Research-Experience Employee	35,000	32,000	5,400	700

As Figure 1 showed, the majority of Startups fail within 5-years and more than 90% of Startups either die or hire fewer than 10 employees within the first 5-years of existence. We also pay special attention to high growth startups and "industrial" startups (defined as being a startup engaged in either manufacturing, information technology, finance, professional/scientific services and health care. Figure 3 shows the size distribution for all startups, along with their average earnings distribution at time *t=0*.

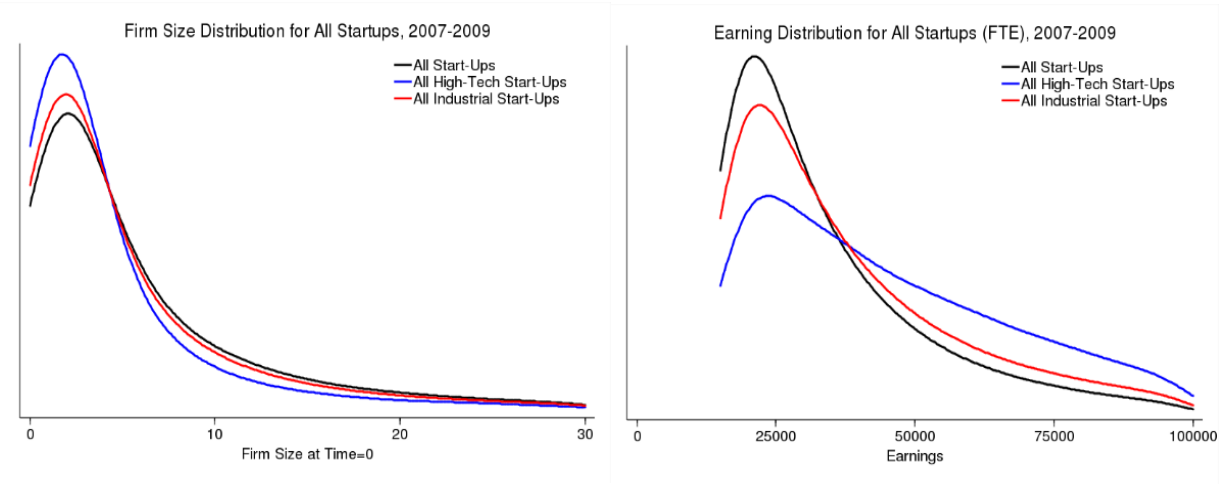


Figure 3: Startup Size and Earnings Distribution at time *t=0*

The vast majority of startups are extremely small in their first year as 75% of all startups have fewer than 5 employees at time *t=0*, with more than 50% of startups having 2 or fewer employees. Fewer than 5% of Startups hire more than 20 employees in the initial period. This is consistent across all startup types as well. Most startups pay relatively small earnings, with startups in High-tech industries typically offering the highest earnings.[11] These two findings, combined with the high-rate of failure suggest that startups face significant capital constraints. The small size also highlights the importance of human capital in the initial period.

Human capital composition

Table 6 below provides the total number of startup employees, along with the proportion of employees that have R&D-experience, high-tech experience, University experience and research grant experience[12] within the 3-years prior to joining the startup.

[11] The earnings measures do not capture full-year or full-quarter earnings.
[12] Note that about 25,000 of the research-experienced individuals working in startups are directly identified through UMETRICS data. The balance are derived from the machine-learning algorithm

Table 6: Startup Employment Composition[13]

Year	Total ever employed at startup	R&D-experience	High-tech Experience	University Experience	Research-experienced
2006	6.82M				0.09
2007	6.47M				0.09
2008	5.74M				0.09
2009	4.7M	19.3	11.1	2.6	0.09
2010	4.56M	20.3	12	2.2	0.10
2011	4.37M	21.2	13.7	2.4	0.10
2012	4.53M	21.1	13.4	2.6	0.09
2013	4.4M	22.2	14.2	2.7	0.09

Source: Startup Worker History and Startup Firm History Files.

Approximately one in five workers in a startup has experience in an R&D performing firm and one in ten has experience in a High-tech firm. About 3% of the startup workforce is affiliated with a university in the 3-years prior to the startup, and roughly 5% of the university affiliated workforce has worked on a research grant. Table 7 shows the human capital composition by startup type.

Table 7: Human-Capital Composition by Startup Type

	Former High-Tech Employees	Former R&D Employees	Former University Employees
All Startups	10%	17%	2%
High-tech Startups	94%	26%	4%
Industrial Startups	16%	20%	3%

Source: Startup Worker History and Startup Firm History Files.

The table makes it clear that High-tech startups are nearly entirely composed of High-tech employees and have much greater proportions of workers who were previously at R&D performing firms. They also have twice as many former university employees as other startups. Similarly, industrial startups have higher proportions of employees with experience at High-tech and R&D performing firms, as well as more employees with university experience.

Startup Outcomes and workforce characteristics

This section provides some initial descriptive results about the link between workforce experience and startup outcomes.

The outcome variables of interest are measured as follows:

1. Survival to period *t+1*,

[13] Since we focus on the prior three years work experience, the table is left-censored

2. Success (defined as having survived for at least 5-years and employ 10+ employees at time *t+5*),
3. High Growth (defined as having survived for at least 5-years, employ 10+ employees at time *t+5* and be in top ten percentile of employment growth among your cohort (conditional on employing 5+ employees at time *t=0*)),
4. Employment Growth to *t+1* (conditional on having at least 5+ employees at time *t=0*),
5. Employment Growth to *t+5* (conditional on having at least 5+ employees at time *t=0*).

We standardize our descriptive analysis by defining a startup's workforce as "intensive" in one of our human capital dimensions if it employs more of a certain type of worker than the median startup within a size group. This means that for all startups of size ten or more employees at time t=0 for example, we compare the outcomes of startups that employ disproportionately more R&D workers to startups that employ disproportionately less R&D workers. The results for survival outcomes are reported below in Figure 4.

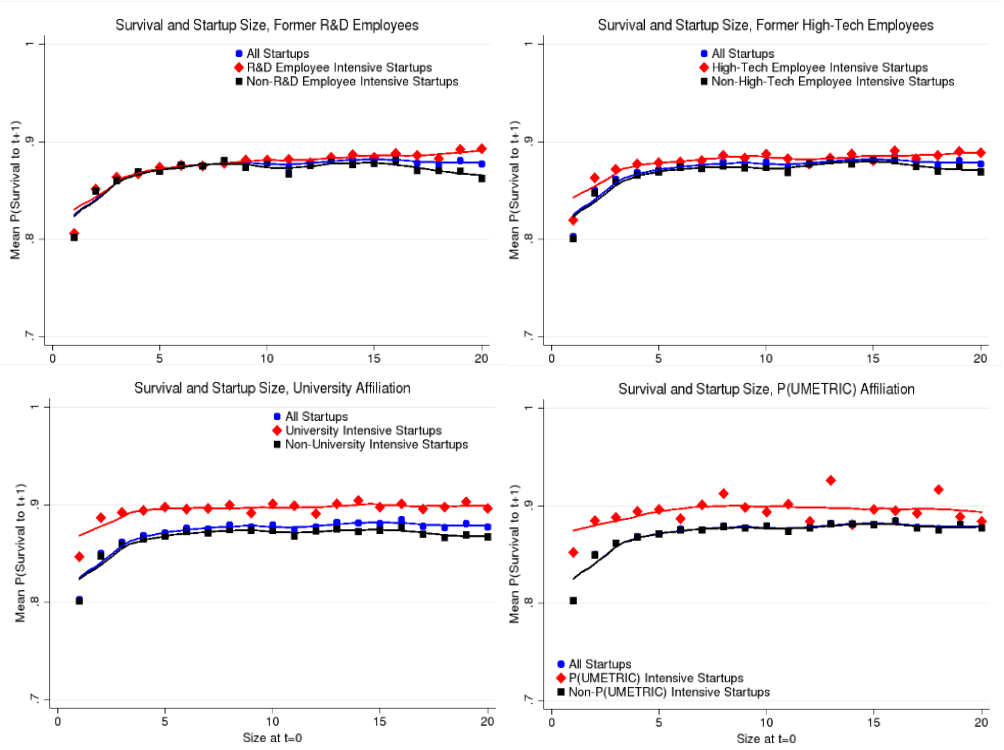


Figure 4: Survival by Human capital intensity

Figure 4 is consistent with the view that startups with higher proportions of high human capital employees are more likely to survive. The slope rises for very small startups and then quickly flattens out, indicating that the probability of survival to the next period does not change once a startup is larger than five employees at time t=0. We see a clear separation in the survival probabilities of startups that hire University employees and research trained (UMETRICS) employees intensely. There is minor separation in the survival probabilities for High-tech startups and almost no difference in the survival probabilities between employees with and without experience in R&D performing firms.

Figure 5 shows the results of a similar analysis using a measure of whether the startup was successful (defined as having 10+ employees and surviving for 5+ years).

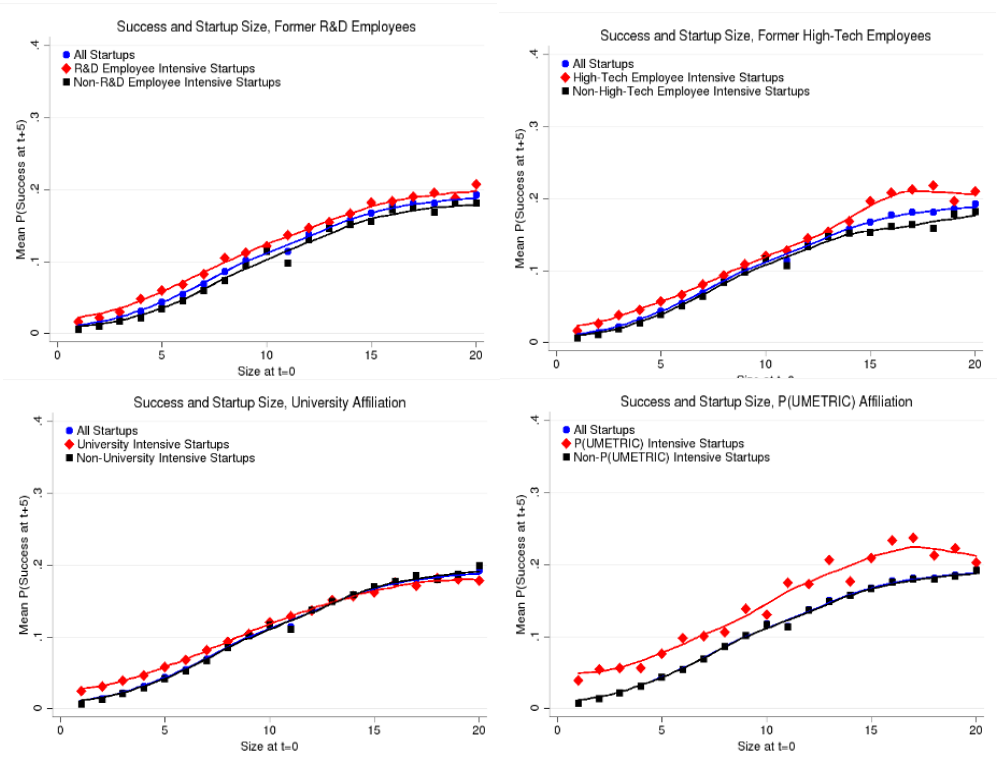


Figure 5: Startup Success and Human Capital Intensity

The figures are all upward sloping, indicating that the likelihood of being defined as a success rises as the initial size of the startup increases.There are minor differences in the probability of startup success for those startups that hire R&D employees intensively, as well as startups that hire High-tech employees intensively. Interestingly, there is almost no difference in the success outcomes for startups that hire university employees intensively. However, there is a substantial difference in the success outcomes for startups that hire research-trained (UMETRICS) employees.

Finally, Figure 6 shows the results of a similar exercise that compares the outcomes for whether a startup is a high growth startup (defined as having 10+ employees and being in the top 10% of the employment growth rate distribution within their startup year cohort).

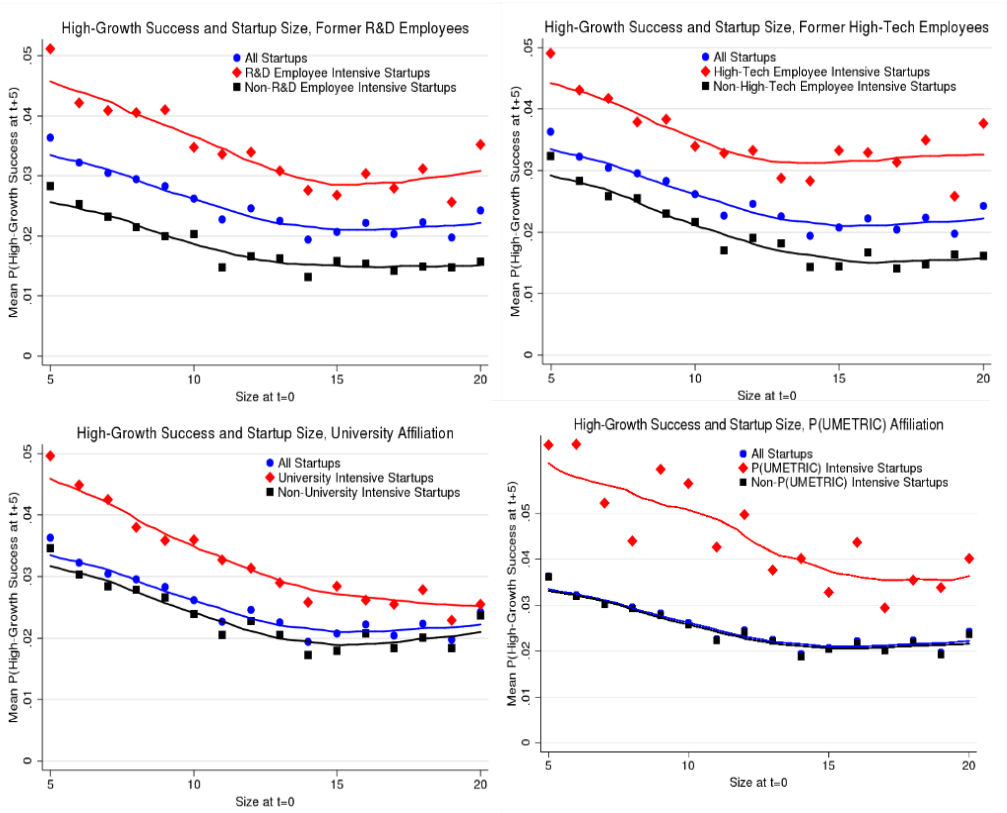


Figure 6: High-Growth Success and Human Capital Intensity

Here, the figures are all showing a downward sloping line, indicating that the likelihood of being considered a high-growth successful startup declines as the initial size of the startup increases. This is likely due to construction as the potential for higher rates of growth increases for firms that initially start out as smaller. The shape of the trend is less important than the separation that exists between intensity measures. There are clear differences in the probability of high growth success across all designations of human capital. The results are consistent with those shown in Figures 3 and 4, but do display more dispersion. This may be due to the fact that these high growth firms make up fewer than 1% of the total number of startups, creating more volatility and disclosure restrictions for the subset of firms that hire more than 20 employees in the initial period.

Analytical Results

The basic framework was provided in Equation (1). We assume that the functional form of Equation (1) is a linear combination of exponential functions, allowing us to use a log-linear estimation and calculate multiple outcome measures for each startup (survival, “success”, “high-growth success” and employment growth) both one and five years after the birth of the firm. We regress these outcomes against the startup’s workforce and other characteristics in the year of firm birth (t=0).

Our main empirical specification is as follows

$$Y_f = \alpha + \beta_1 \ln EARN_{f0} + \sum_{k=1}^{9} \delta_k SIZE_{kf0} + \beta_2 \ln \overline{AGE_{f0}} + \beta_3 \ln FEMALE_{f0} + \beta_4 \ln FOREIGN_{f0} + \beta_5 \ln RD_{f0} + \beta_6 \ln HT_{f0} + \beta_7 \ln UNI_{f0} + \beta_8 \ln Research\ Experience_{f0} + \varepsilon$$

The key measures of interest are the workforce human capital measures – the number of workers who have worked in R&D performing firms, High-tech firms, universities – as well as the number who have direct research experience. Since the Census Bureau does not have direct measures of technology, we control for industry, detailed geography and year using fixed effects. We also include mean earnings of the workforce as well as firm employment size categories. External macroeconomic conditions are proxied by zip code-year fixed effects and industry fixed effects. We interact demographics with each of the R&D worker types to identify potential non-linearities of being a certain type of worker (e.g. female University worker).[14] Due to the fact that majority of human capital measures will be zero, rather, than using a log transformation, we instead transform each of the coefficients to an inverse hyperbolic sine in order to minimize the selection biases from dropping startups with zeros in the human capital categories, so that for each variable, we have:

$$\ln(x) \rightarrow \log(x + (x^2 + 1)^{1/2})$$

The first specification separates all of the human capital designations in order to separately describe the relationship between each type of human capital and startup outcomes before applying control factors. We assess outcome measures relating to whether or not the startup becomes a "success" or whether it can be classified as a "high-growth" startup. The second specification will identify how the human capital variables impact startup growth-rates.

Figure 7 reports the coefficient estimates of the standalone human capital designations by firm-size for two separate outcomes: survival and success. The results show that the standalone human capital coefficients decline as the firm gets bigger, highlighting that there may be diminishing marginal returns to the employment of each additional type of worker. The returns to each type of worker declines very rapidly for the survival outcome, with a more modest and steady decline in the success rates.

[14] Note that these interaction terms are the result of multiplying continuous counts of employees falling into each group and that any given employee may belong to any number of designated groups.

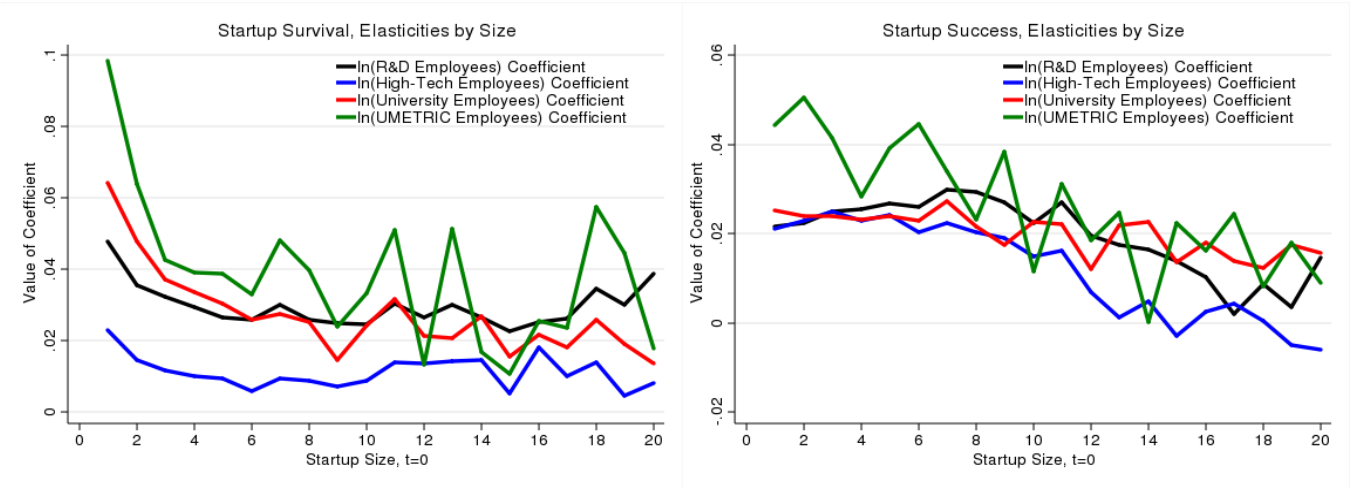


Figure 7: Coefficient Values of Standalone Human Capital Measures by Firm Size

Table 8 provides the key results associated with the full regression (tables that include all control variables can be found in the Appendix). Briefly, all measures of workforce R&D-experience are positively and significantly related with startup success and growth, with former R&D and high-tech employees having a negative effect on first year survival. The coefficient for *research experience* is additive with the *university* coefficient as all research-experienced employees are also former university employees. A one-unit increase in this worker-type leads to approximately 1.86 percentage point increase in the survival rate to year t+1, a 2.7 percentage point increase in the probability of becoming a successful startup and a 1.67 percentage point increase in becoming a high-growth. Considering that the probability of becoming a successful startup is 6.6%, the benefit of adding one additional worker with research experience increases the likelihood of becoming a successful startup by more than 40%. The probability of becoming a high-growth (employment-based) successful startup is 1.2%, so that the benefit of adding an additional research-experienced worker more than doubles.

In terms of becoming a high growth (revenue-based) startup, having prior experience at an R&D firm, high-tech firm or university has a positive and significant impact. Having direct research experience has a positive impact, but is weakly significant. As a proxy for productivity, we look at the probability of becoming a high growth startup based on growth to revenue per employee. We see that the human capital elements have either a weakly positive or significantly negative effect on the likelihood of a startup being one of the fastest growers based on revenue per employee growth. Finally, to summarize, hiring each type of R&D worker has a positive effect and the interaction variables are also positive and significant. (see Appendix for details).

Table 8: OLS on All Startup Outcomes, 2005-2014

Outcome Variable	Survival, year 1	Success, year 5	High Growth (Employment-based), year 5	High Growth: (Revenue-based), year 5	High Growth: (Revenue per Employee-based), year 5
ln RD_{f0}	-0.00200***	0.0260***	0.0107***	0.00256***	-0.00258***
	(0.000372)	(0.000472)	(0.000222)	(0.000283)	(0.000279)
ln HT_{f0}	-0.00893***	0.0198***	0.00790***	0.00600***	-0.00117**
	(0.000486)	(0.000616)	(0.000289)	(0.000369)	(0.000364)
ln UNI_{f0}	0.00433***	0.0220***	0.00866***	0.00274***	0.000258
	(0.000803)	(0.00111)	(0.000521)	(0.000666)	(0.000656)
ln $research\ experience_{f0}$	0.0143***	0.00578	0.00809***	0.00306	0.000259
	(0.00299)	(0.00367)	(0.00172)	(0.00220)	(0.00217)
Zip Code-Year FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Observations	3,730,000	3,730,000	3,730,000	3,730,000	3,730,000
R-squared	0.079	0.193	0.049	0.018	0.014

Robust Standard Errors in Parentheses. *p<0.05, **p<0.01, ***p<0.001; controls included for size and average earnings, proportion of workforce that is female, foreign born, and interactions of female, foreign born with research experience. Full results in the appendix

Table 9 looks at the impact of the same coefficients on the growth rates for employment, revenue and revenue per employee.

Table 9: OLS on All Startup Growth Rates, 2005-2014

Outcome Variable	Employment Growth, t+1	Revenue Growth, t+1	Revenue per Employee Growth, t+1
ln RD_{f0}	0.151***	-0.0147***	-0.107***
	(0.00117)	(0.00139)	(0.00157)
ln HT_{f0}	0.0483***	0.00636***	-0.0501***
	(0.00145)	(0.00177)	(0.00201)
ln UNI_{f0}	0.0517***	0.0232***	-0.0347***
	(0.00219)	(0.00307)	(0.00348)
ln $research\ experience_{f0}$	0.0177*	0.00109	-0.0131
	(0.00739)	(0.0110)	(0.0124)
Zip Code-Year FE	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes
Observations	3,730,000	3,730,000	3,730,000
R-squared	0.079	0.193	0.049

Robust Standard Errors in Parentheses. *p<0.05, **p<0.01, ***p<0.001; controls included for size and average earnings, proportion of workforce that is female, foreign born, and interactions of female, foreign born with research experience. Full results in the appendix

Table 9 highlights the Davis, Haltiwanger and Schuh (DHS) (2006) growth rates on employment, revenue and revenue per employee in the following year.[15] On the employment growth rate, we find that each human capital component has a positive and significant effect on employment growth in year *t+1*. For revenue growth, we find that startups hiring former high-tech workers and university workers experience higher revenue growth rates. However, the growth rate in revenue fails to keep pace with the growth rate in employment, leading to lower revenue per employee growth. This is suggestive that these startups are converting revenue gains into additional employees.

Tables 10 and 11 report the results for two different categories of startups - industrial startups and high-tech startups. The results in Table 10 are substantively unchanged but there are a few noticeable differences, namely that the research-experienced coefficients are mostly insignificant with the exception of being classified as a high growth startup (either employment or revenue-based). Three of human capital measures have a negative impact on survival to year 2, but all have positive and significant impacts on whether the startup is successful and/or classified as a high-growth (either employment or revenue-based). The impact of these human capital measures on our proxy for productivity growth is less significant and/or negative. The interpretation of the coefficients suggests that hiring one additional research-experienced employee increases the probability of becoming a high-growth success by 3.0 percentage points. This represents an increase of around 250% over the mean (1.6%). The impact of human capital on the growth rates to employment and revenue for industrial startups follows a similar pattern to the growth rates for all startups in terms of signs and significance, with only minor differences in the magnitudes.

Table 10: OLS on Industrial Startup Outcomes, 2005-2014

Outcome Variable	Survival, year 1	Success, year 5	High Growth (Employment-based), year 5	High Growth: (Revenue-based), year 5	High Growth: (Revenue per Employee-based), year 5
ln RD_{f0}	-0.0183***	0.0273***	0.0151***	0.00544***	-0.00218***
	(0.000700)	(0.00105)	(0.000541)	(0.000725)	(0.000640)
ln HT_{f0}	-0.0128***	0.0343***	0.00914***	0.00863***	-0.00336***
	(0.000700)	(0.00100)	(0.000515)	(0.000691)	(0.000610)
ln UNI_{f0}	-0.00408***	0.0289***	0.0123***	0.00730***	0.000931
	(0.00124)	(0.00208)	(0.00107)	(0.00143)	(0.00127)
ln $research\ experience_{f0}$	0.00604	0.00578	0.0177***	0.0127*	0.00596
	(0.00474)	(0.00738)	(0.00379)	(0.00508)	(0.00448)
Zip Code-Year FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Observations	1,134,000	1,134,000	1,134,000	1,134,000	1,134,000
R-squared	0.069	0.213	0.079	0.306	0.203

Robust Standard Errors in Parentheses. *p<0.05, **p<0.01, ***p<0.001; controls included for size and average earnings, proportion of workforce that is female, foreign born, and interactions of female, foreign born with research experience. Full results in the appendix

[15] The growth rate formula is given by: $GROWTH_{it+1} = 2 \times \left(\frac{SIZE_{it+1} - SIZE_{it}}{SIZE_{it+1} + SIZE_{it}}\right)$

Turning now to high-tech startups in Table 11, we find that the impact of hiring former university and research-experienced workers is mostly insignificant, while the impact of hiring high-tech employees at year 0 has significant and positive effects on success, high-growth success and employment growth.

Table 11: OLS on High-tech Startup Outcomes, 2005-2014

Outcome Variable	Survival, year 1	Success, year 5	High Growth (Employment-based), year 5	High Growth: (Revenue-based), year 5	High Growth: (Revenue per Employee-based), year 5
ln RD_{f0}	-0.00704***	0.0308***	0.0173***	0.0103***	-0.00333*
	(0.00165)	(0.00241)	(0.00127)	(0.00205)	(0.00165)
ln HT_{f0}	-0.0342***	0.0679***	0.0114***	0.00949***	-0.00674***
	(0.00165)	(0.00221)	(0.00116)	(0.00188)	(0.00151)
ln UNI_{f0}	0.00627*	0.00959*	0.00440	0.000488	0.00453
	(0.00271)	(0.00438)	(0.00230)	(0.00372)	(0.00299)
ln $research\ experience_{f0}$	0.00984	0.0185	0.0201**	0.0313*	0.0127
	(0.00968)	(0.0145)	(0.00764)	(0.0124)	(0.00994)
Zip Code-Year FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Observations	148,000	148,000	148,000	148,000	148,000
R-squared	0.112	0.193	0.133	0.536	0.429

Robust Standard Errors in Parentheses. *p<0.05, **p<0.01, ***p<0.001; controls included for size and average earnings, proportion of workforce that is female, foreign born, and interactions of female, foreign born with research experience. Full results in the appendix

In addition to these tables, we have estimated the same specification over different size groups of startups and find that the results are robust and do not differ greatly. To summarize our empirical findings, we find mostly positive and significant associations between R&D-experience (categorized as either having been employed by an R&D performing firm, a high-tech firm, a research university and/or having direct research experience) and startup performance. Startups that hire employees with these human capital measures are more likely to survive, be considered successful and grow faster.

Summary

This paper leverages new data about workforce human capital that can be used to provide more insights into the survival and employment growth of new businesses. These results are consistent with the view that there is a relationship between workforce experience and business startup and survival. Further work using these data will be necessary to examine temporal dynamics. It will be particularly interesting to understand whether changes in the fluidity of this type of workforce, or changes in the nature of research funding, can be tied to the decline in business dynamism.

References

Abowd, John M et al. 2005. “The Relation among Human Capital , Productivity and Market Value: Building up from Micro Evidence.” In *Measuring Capital for the New Economy*, eds. Carol Corrado, John Haltiwanger, and Sichel Dan. University of Chicago Press, 153–204.

———. 2009. “The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators.” In *Producer Dynamics: New Evidence from Micro Data*, University of Chicago Press, 149–230.

Abowd, John M, John Haltiwanger, and Julia Lane. 2004. “Integrated Longitudinal Employer-Employee Data for the United States.” *American Economic Review* 94(2): 224–29. http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=13708218&site=ehost-live&scope=site.

Acemoglu, Daron, Ufuk Akcigit, Nicholas Bloom, and William R Kerr. 2013. *Innovation, Reallocation and Growth*. National Bureau of Economic Research.

Andrews, Dan, Chiara Criscuolo, and Peter N. Gal. 2015. *Frontier Firms, Technology Diffusion and Public Policy*. OECD Publishing.

Bania, Neil, Randall W Eberts, and Michael S Fogarty. 1993. “Universities and the Startup of New Companies: Can We Generalize from Route 128 and Silicon Valley?” *The review of economics and statistics*: 761–66.

Barth, Erling, James Davis, and Richard B Freeman. 2016. “Augmenting the Human Capital Earnings Equation with Measures of Where People Work.” *National Bureau of Economic Research Working Paper Series* No. 22512.

Bloom, Nicholas, Charles I Jones, John Van Reenen, and Michael Webb. 2016. *Are Ideas Getting Harder to Find*. https://web.stanford.edu/~chadj/IdeaPF.pdf.

Breiman, Leo. 2001. “Random Forests.” *Machine-learning* 45(1): 5–32.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and Regression Trees*. CRC press.

Davis, Steven, and John Haltiwanger. 2014. *Labor Market Fluidity and Economic Performance*. Cambridge, MA. http://www.nber.org/papers/w20479.pdf (February 19, 2017).

Decker, Ryan A., John C. Haltiwanger, Ron S. Jarmin, and Javier Miranda. 2014. “The Role of Entrepreneurship in US Job Creation and Economic Dynamism †.” *Journal of Economic Perspectives* 28(3): 3–24. http://pubs.aeaweb.org/doi/abs/10.1257/jep.28.3.3.

———. 2016a. *Declining Business Dynamism: Implications for Productivity?* https://www.brookings.edu/research/declining-business-dynamism-implications-for-productivity/ (February 17, 2017).

———. 2016b. “Declining Business Dynamism: What We Know and the Way Forward †.” *American Economic Review* 106(5): 203–7. https://www.aeaweb.org/articles?id=10.1257/aer.p20161050 (June 2, 2016).

———. 2016c. “Where Has All the Skewness Gone? The Decline in High-Growth (Young) Firms in the U.S.” *European Economic Review* 86: 4–23. http://www.sciencedirect.com/science/article/pii/S0014292116300125 (May 27, 2016).

———. 2017. *Changing Business Dynamism and Productivity: Shocks vs. Responsiveness*.

Fan, Rong-En et al. 2008. “LIBLINEAR: A Library for Large Linear Classification.” *Journal of machine-learning research* 9(Aug): 1871–74.

Fernald, John. 2014. *Productivity and Potential Output Before, During, and After the Great Recession*. Cambridge, MA. http://www.nber.org/papers/w20248.pdf (August 14, 2016).

Fleming, Lee, I I I Charles King, and Adam Juda. 2007. “Small Worlds and Regional Innovation.” *Organization Science* 18(6): 938–54.

Glaeser, Edward L, William R Kerr, and Giacomo A M Ponzetto. 2010. “Clusters of Entrepreneurship.” *Journal of Urban Economics* 67(1): 150–68.

Golan, Amos, Julia Lane, and Erika McEntarfer. 2007. “The Dynamics of Worker Reallocation within and across Industries.” *Economica* 74(293): 1–20.

Goldschlag, Nathan, and Javier Miranda. 2016. “Business Dynamics Statistics of High-tech Industries.”

Goldschlag, Nathan, and Alexander T. Tabarrok. 2014. “Is Regulation to Blame for the Decline in American Entrepreneurship?” *SSRN Electronic Journal*. http://www.ssrn.com/abstract=2559803 (February 19, 2017).

Gordon, Robert J. 2016. 1 *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War*. http://econpapers.repec.org/RePEc:pup:pbooks:10544 (August 14, 2016).

Gutiérrez, Germán, and Thomas Philippon. 2016. *Investment-Less Growth: An Empirical Investigation*. Cambridge, MA. http://www.nber.org/papers/w22897.pdf (February 20, 2017).

Guyon, Isabelle, and André Elisseeff. 2003. “An Introduction to Variable and Feature Selection.” *Journal of machine-learning research* 3(Mar): 1157–82.

Hathaway, Ian, and Robert E Litan. 2014. “Declining Business Dynamism in the United States: A Look at States and Metros.” *Brookings Institution*.

Hausman, Naomi. 2012. *University Innovation, Local Economic Growth, and Entrepreneurship*.

Hecker, Daniel E. 2005. “High-Technology Employment: A NAICS-Based Update.” *Monthly Lab. Rev.* 128: 57.

Hopenhayn, Hugo A. 1992. “Entry, Exit, and Firm Dynamics in Long Run Equilibrium.” *Econometrica* 60(5): 1127. http://www.jstor.org/stable/2951541?origin=crossref (February 19, 2017).

Hopenhayn, Hugo, and Richard Rogerson. 1993. “Job Turnover and Policy Evaluation: A General Equilibrium Analysis.” *Journal of Political Economy* 101(5): 915–38. http://www.journals.uchicago.edu/doi/10.1086/261909 (February 19, 2017).

Hyatt, Henry R, and James R Spletzer. 2013. “The Recent Decline in Employment Dynamics.” *IZA Journal of Labor Economics* 2(1): 5. http://izajole.springeropen.com/articles/10.1186/2193-8997-2-5 (February 19, 2017).

James, Gareth, D Witten, T Hastie, and R Tibshirani. 2013. “An Introduction to Statistical Learning (Vol. 103).”

Jones, Charles I. 2002. “Sources of US Economic Growth in a World of Ideas.” *The American Economic Review* 92(1): 220–39.

Kantor, Shawn, and Alexander Whalley. 2013. “Knowledge Spillovers from Research Universities: Evidence from Endowment Value Shocks.” *Review of Economics and Statistics* 96(1): 171–88.

———. 2014. “Research Proximity and Productivity: Long-Term Evidence from Agriculture.” *Review of Economics and Statistics. Forthcoming*.

Karahan, Fatih, Benjamin Pugsley, and Aysegül Sahin. 2015. “Understanding the 30-Year Decline in the Startup Rate: A General Equilibrium Approach.” *Unpublished manuscript, May*.

Kohavi, Ron, and George H John. 1997. “Wrappers for Feature Subset Selection.” *Artificial*

intelligence 97(1): 273–324.
Lane, Julia I., Jason Owen-Smith, Rebecca F. Rosen, and Bruce A. Weinberg. 2015. “New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value.” *Research Policy*.
Lane, Julia, Jason Owen-Smith, Rebecca Rosen, and Bruce Weinberg. 2014. Research Policy *New Linked Data on Science Investments, the Scientific Workforce and the Economic and Scientific Results of Science*.
Lowe, Robert A, and Claudia Gonzalez-Brambila. 2007. “Faculty Entrepreneurs and Research Productivity.” *The Journal of Technology Transfer* 32(3): 173–94.
Marx, Matt, Jasjit Singh, and Lee Fleming. 2015. “Regional Disadvantage? Employee Non-Compete Agreements and Brain Drain.” *Research Policy* 44(2): 394–404.
Zhang, Cha, and Yunqian Ma. 2012. *Ensemble Machine-learning*. Springer.
Zolas, Nikolas et al. 2015. “Wrapping It up in a Person: Examining Employment and Earnings Outcomes for Ph.D. Recipients.” *Science* 350(6266): 1367–71. http://www.sciencemag.org/content/350/6266/1367.abstract (December 10, 2015).

Appendix I – Data Construction

IA Startup Firm and Startup Worker History File Data Construction

This section describes the construction of a panel data set containing the full firm and worker history of all employees affiliated with startups. We start by first describing the construction of the startup firm history file before delving into the details on the construction of the worker history file.

IA.1 Startup Firm History File, 2005-2014

The Startup Firm History File is based entirely off the Longitudinal Business Database (LBD) (see Jarmin and Miranda 2002 for details on construction), a longitudinally linked establishment-level dataset that allows users to identify firm (and establishment) births and deaths. The database contains the universe of private non-farm businesses in the United States beginning from 1976 until 2014 and contains variables such as employment, industry (NAICS or SIC), country-level geographic identifiers, payroll, legal form of organization and more.

Our human capital measures will be derived from two databases housed at Census that include W2's (beginning from 2005 to 2014) and LEHD (variation in start dates by state, but widespread coverage begins around 2000). Because our main individual-level measures will begin in 2005, we limit the startup cohorts to start from 2005 and later. Our Startup Firm History File simply keeps firms whose birth dates (first appearance in the LBD) occurs from 2005 and afterwards. Table A1 highlights our starting frame and our full startup frame.

Year	LBD Firm Count	New Firms by Year Cohort	Firms Born After 2005
2005	6,359,000	648,000	648,000
2006	6,350,000	673,000	1,132,000
2007	6,365,000	654,000	1,523,000
2008	6,265,000	584,000	1,788,000
2009	6,090,000	490,000	1,924,000
2010	6,042,000	470,000	2,083,000
2011	5,999,000	489,000	2,254,000
2012	6,051,000	531,000	2,477,000
2013	6,103,000	746,000	2,922,000
Total	55,624,000	5,285,000	16,751,000

IA.2 Startup Worker History File, 2005-2014

Our analysis focuses on the human capital composition of the startups at their inception (birth year) to assess how this composition can predict future outcomes. Using the Startup Firm History File and linked employer-employee data, we can easily identify the workers for startups at time *t=0*. We will then track their prior and post-startup earnings history and experience. Our final dataset will be organized by PIK-EIN-Year and contain earnings history data, along with EIN-level characteristics such as size, payroll, industry and geography.

Our construction starts with the W2 as the frame of the database. We use the W2 and not the LEHD due to coverage issues associated with the LEHD, including missing state-years, missing university work-history for work-study, missing self-employed and misreporting by firms to state UI. There is also the issue of reporting results under the LEHD program.

We supplement the W2 with unmatched LEHD observations (i.e. observations captured in LEHD, but not in W2) and with LEHD geocodes in order to generate better match rates to the establishments, but our earnings history will be primarily derived from W2.

1. Combine all of the years of the W2 into one master file organized by PIK-EIN-Year. Keep the "wages_tip" variable for your earnings and drop the "fica_wage" variable as the "wages_tip" appears to be more comprehensive and complete. Master W2 file is sorted by PIK-EIN-Year and contains 2,281M observations from 2005 until 2014.

2. Combine the Employee History File (EHF) and Employer Characteristics File (ECF) into one master file so that the EHF file is organized by PIK-EIN-Geocode-Year.
    a. To do this, we first start with annualized data of the full EHF file that is organized by PIK-SEIN-SEIN UNIT (first implicate). We construct annual measures of earnings by combining the quarterly data (organized as e101, e102, etc…) in the raw file. The annualized EHF file is organized by PIK-SEIN-SEINUNIT-Year-Earnings. It contains 2,105M observations starting from 2005 until 2014.
    b. Create annualized ECF file. The annualized ECF file contains SEIN-SEINUNIT-Year-EIN-Geocode. The data is originally organized by quarter in the raw files, so to construct the annualized version, I only keep the five variables specified above and drop ALL DUPLICATES of these 5 variables. Data is organized by SEIN-SEINUNIT-Year-EIN-Geocode with 98.8M observations starting from 2005 until 2014.
    c. Combine ECF and EHF file into one file organized as PIK-EIN-Geocode-Year-Earnings. I first sort the two components by SEIN-SEINUNIT-Year. Once merged, I drop the SEIN and SEINUNIT variables and just leave the PIK-EIN-Geocode-Year-Earnings combined file. I drop all of the observations with missing PIK or missing Earnings. By EHF counts, we get very high match rates (~97% overall), but by ECF counts, we only get a match rate of 83%.

Match Rate Statistics for Combined ECF-EHF file:

Year	EHF Observations	Match EHF Observations	EHF Match Rate	ECF Observations	Matched ECF Observations	ECF Match Rate
2005	216.6M	209.4M	96.7%	9.5M	7.9M	83.2%
2006	222M	214.7M	96.7%	9.7M	8.1M	83.3%
2007	223.1M	216M	96.8%	9.8M	8.2M	83.2%
2008	214.2M	207.5M	96.9%	10M	8.2M	82.6%
2009	192.8M	186.9M	96.9%	9.9M	8.1M	82.3%
2010	197.3M	191.7M	97.1%	9.8M	8.2M	84.1%
2011	201.9M	196.6M	97.3%	9.9M	8.3M	84.3%
2012	207.1M	202.2M	97.7%	10M	8.4M	84.2%
2013	213M	208.7M	98.0%	10M	8.5M	85.2%
2014	217.7M	216.4M	99.4%	10.2M	8.7M	84.9%
Overall	2105.7M	2050.1M	97.4%	98.8M	82.6M	83.6%

We can see that the EHF match rate is consistently high and above 95%, while the ECF match rate is in the low 80%

3. I then merge the Combined EHF-ECF File (2,050M observations) to the master W2 file constructed in Step 1 (2,281M observations). To do this, I match each dataset by PIK-EIN-Year. The unmatched data on each side of the file are kept in the master file. The combined file contains PIK-EIN-Geocode-Year-W2 Earn-LEHD Earn.

Employee-Match Rate Reporting Statistics:

Year	W2 Observations	Match W2 Observations	W2 Match Rate	EHF-ECF Observations	Matched EHF-ECF Observations	EHF-ECF Match Rate
2005	237.9M	187.2M	78.70%	209.4M	187.2M	89.40%
2006	243.1M	191.8M	78.90%	214.7M	191.8M	89.30%
2007	244M	193.4M	79.20%	216M	193.4M	89.50%
2008	235.2M	186.2M	79.20%	207.5M	186.2M	89.70%
2009	214.3M	168.2M	78.50%	186.9M	168.2M	90.00%
2010	208.7M	168.2M	80.60%	191.7M	168.2M	87.80%
2011	212.7M	173.4M	81.50%	196.6M	173.4M	88.20%
2012	223.6M	178.5M	79.80%	202.2M	178.5M	88.30%
2013	229.2M	189.1M	82.50%	208.7M	189.1M	90.60%
2014	232.3M	192M	82.70%	216.4M	192M	88.70%
Overall	2280.9M	1828M	80.10%	2050.2M	1828M	89.16%

The match rate is around 80% for the W2, meaning that LEHD is missing approximately 20% of W2 observations. Meanwhile, the match rate on the LEHD side is around 90%, meaning that approximately 10% of W2 observations are not found in LEHD. The next table combines and merges the matched and unmatched pairs to generate a complete universe of PIK-EIN-Year combinations.

Combined W2-EHF-ECF File Observations

Year	Combined EHF-ECF-W2 Observations	W2 Observations	Additional Observations from LEHD	EHF-ECF Observations	Additional Observations from W2
2005	271.8M	237.9M	33.9M	209.4M	62.4M
2006	278M	243.1M	34.9M	214.7M	63.3M
2007	278.6M	244M	34.6M	216M	62.6M
2008	267.9M	235.2M	32.8M	207.5M	60.4M
2009	242.9M	214.3M	28.6M	186.9M	56M
2010	241.8M	208.7M	33M	191.7M	50.1M
2011	245.4M	212.7M	32.7M	196.6M	48.8M
2012	256.5M	223.6M	32.9M	202.2M	54.2M
2013	257.9M	229.2M	28.7M	208.7M	49.1M
2014	262.9M	232.3M	30.6M	216.4M	46.5M
Overall	2603.7M	2280.9M	322.8M	2050.2M	553.5M

The full universe of employee-employer matches that we will be checking whether they worked for startups consists of more than 2,604M observations, of which 1,828M are found in both the W2 and LEHD, 323M are found only in the LEHD and not W2 and 554M are found only in the W2 but not LEHD.

4. The database here consists of PIK-EIN-Year-Earnings-Geocode. The next step involves matching the full database to a set of establishment variables. In order to do this, we first need to construct a panel database of establishment attributes. Thankfully, a database is already in existence that contains this! We start with the Longitudinal Business Database (LBD) as our frame, compiled from 2005 until 2014. This database contains an establishment identifier (LBDNUM), FirmID, Industry code, Employment and Payroll. It also contains data on the entry and exit date of each firm, which we will use to identify startups.

   Missing from the LBD, is the EIN code for each establishment and geographic identifier. We fill this gap in the data by merging in the Business Register, which contains both of these items.

Our matching algorithm then links the combined W2/LEHD file to the BR/LBD file by the following criteria (sorted from best to worst).

1. EIN & 11-digit GEOCODE & YEAR -
2. EIN & 11-digit GEOCODE
3. EIN & 5-digit GEOCODE & YEAR
4. EIN & 5-digit GEOCODE
5. EIN & 2-digit GEOCODE & YEAR
6. EIN & 2-digit GEOCODE
7. EIN & YEAR
8. EIN

Once completed, we have establishment information linked to each of the employees, which will allow us to assign industry descriptors such as whether or not the employee was linked to a high-tech firm. This completes the construction of the Startup Worker History File. The next section outlines the construction of the dataset used in the machine-learning exercise.

IB Machine Learning Training Data Construction

The objective of this exercise is to classify individuals paid by top research universities as to whether or not they are likely to have participated in grant funded research. Our strategy is to leverage the information found in the UMETRICS data to train a supervised machine learning model. We leverage a number of data sources to create a rich set of features to predict research status, including individual demographic characteristics, individual employment history, and university and firm characteristics.

The at-risk set includes all individuals observed in the W2 data paid by a corresponding university EIN between 2005 and 2014. The university EINS are derived from multiple sources including IPEDS (source of the primary frame), NSF and NIH federal research outlays, and IRS non-profit directory. Our training data includes 14 UMETRICS institutions for which we observe both individuals that have participated in grant funded research (UMETRICS data) and those who have not (defined as those linked to the universities in the W2 during the same period but not found in the UMETRICS data).

Our methodology proceeds in the following steps. First, we create a training set that will be used to inform the classification models. Second, we create the target set (test set or out-of-sample set) on which the classification model will predict whether individuals participated in grant funded research. Third, we gather a rich set of person and institution level features (attributes or characteristics) that will be used to separate individuals in the classification models. Fourth, we perform a series of feature selection exercises to avoid over-fitting, reduce computational burden, and improve prediction quality. Fifth, we estimate several classification models including logistic regression, decision tree, and random forest. Finally, we execute a number of cross validation exercises to test the quality and robustness of the model predictions.[1]

IB.1 Training Data Construction

To build the training set we integrate UMETRICS and W2 data. We use UMETRICS data to classify all individuals associated with UMETRICS institutions in the W2 data as to whether they participated in grant funded research (true positives) and those who have not (true negatives).[2] The training set is derived from the November 2016 vintage of the UMETRICS data, which includes 16 institutions. These institutions alone account for about 20% of federally funded university research spending in 2014. We exclude Stony Brook, and University of Hawaii due to data quality issues[3]. These issues include problems with the employee ID taken from IRIS where running zeroes are dropped and we have lower PIK rates than in other schools (case for Hawaii). Other issues include the fact that the EIN code for some of these schools are not broken out for each satellite school in the university system, meaning that the number of zeroes is extraordinarily high

[1] All feature selection and learning algorithms are drawn from Scikit Learn Python libraries (Scikit-learn 2011).
[2] The period for which these true positives and true negatives are known is dependent on the institution. These true values are defined in years of overlap between the UMETRICS data and the W2 from 2005-2014.
3 Hawaii's EIN code only marginally shows up in the W2s. We need to allocate the proper EIN code, which will require further iterations. Stony Brook, on the other hand has a glut of observations due to all of the satellites being listed under the same EIN code. Therefore, it is impossible to be certain whether an individual affiliated with a SUNY school is actually a UM=1 or UM=0.

as a proportion of UMETRIC employees. The key dilemma faced in this section is that we want to ensure that the zeroes (e.g. unmatched employees of UMETRIC universities) are actually zeroes, otherwise they will contaminate our training data set.

To draw out our target set, we use the EIN codes affiliated with the UMETRIC institutions and merge them directly with the W2 data to identify all of the individuals affiliated with the UMETRIC institutions. We then take those individuals and match them again to the W2 data to draw out their earnings history, prior to being affiliated with the UMETRIC institution and after. This gives us a panel data set of PIK-EIN-Year-Earnings for all individuals affiliated with UMETRIC institutions.

We then take the EIN data and merge them to Census business data by going through a 9-step merging process. We first match the PIK-EIN-Year combinations to the LEHD data, which contains some geographic attributes associated with PIK-EIN-Year combinations, specifically, an 11-digit geocode identifier. The 11-digit identifier is found on approximately 78% of observations and is used to specifically identify the establishment of the PIK-EIN, in order to generate industry and geographic characteristics of the firm that we can compare against. Once we have the Geocodes, we implement the following match program to the LBD and BR/SSEL. The LBD contains the establishment-level data we are interested in, namely employment, payroll, industry (6-digit NAICS), a 5-digit GEOCODE (State and FIPS), LFO and more. We supplement the LBD with the BR/SSEL, which contains EIN code and an 11-digit GEOCODE (in many instances).

For the initial matching program, we start with: 6.8M observations of employees who at one point had been affiliated with a university and their full work history. We perform the matching to include all establishment and firm-level characteristics associated with the employment history of the individual.

1. EIN & 11-digit GEOCODE & YEAR -
2. EIN & 11-digit GEOCODE
3. EIN & 5-digit GEOCODE & YEAR
4. EIN & 5-digit GEOCODE
5. EIN & 2-digit GEOCODE & YEAR
6. EIN & 2-digit GEOCODE
7. EIN & YEAR
8. EIN

The matched dataset contains the same number of PIK-EIN-Year observations as the initial frame, along with other variables collected from the LBD and BR. These include: EMP, PAY, AGE, NAICS, GEOCODE, ZIP.

The next task involves converting the long file with the associated variables and generating a bunch of new variables that look at the Pre/Post earnings history of the individual (from when they entered and exited the university). We do this in a number of ways. We first take the first year and last year that the individual entered and exited the university. We then can generate variables on the 2-digit industry code that the individual worked in (before, after and during), the state (in state/out-state, all 50 states) the individual worked in before, after and during, the characteristics

of the dominant employer (before, after and during) which would include size, payroll, average earnings, industry, location, and additional earnings information such as earnings from other jobs and so forth. We do this for the period t-2 until t+2 for the individual entering and exiting the university. We also separate out the variables by year. The final result is a PIK-UniversityEIN level training database that contains more than 1,300+ unique variables to compare against.

Once the Census side of the training database is complete, we incorporate a number of university-level characteristics gathered from the Carnegie Institute, including enrollment size, average SAT score, indicators for whether or not the university has a medical school, whether it is public/private institution, land grant university and more and assign them to each Individual-UniversityEIN combination. We also include total federal outlays from the NSF by university (collected from NCSES[4]) and total federal outlays from NIH by university and number of NIH awards. These are collected for the time period from 2005 until 2014. We also include UMETRICS information such as the years the individual is listed on a grant into the database. This is a manually created dataset combining university data from multiple sources and arranged across multiple EIN classifications for different universities.

Finally, we merge in the Individual Characteristics File from the Decennial into the final dataset. The final training database contains nearly 1,400 variables including demographic data from the ICF, employment history data from the Census and university data collected from NSF, NIH and UMETRICS. There are approximately 1.5M individuals classified as known UMETRICS, with approximately 140,000 of them classified as UMETRICS=1 and the remainder being unclassified. This gives us a very robust training data set to compare pre/post and during university outcomes.

Training Data Issues: For approximately 30,000 UMETRIC individuals, the university EIN listed on UMETRICS differ from the university EIN listed on the W2's. We believe that this may due to the individual being listed on the grant, but affiliated with a different university (as is common on grants). In this case, we identified that individual as being part of the W2 university and drop them from the Training Data.

IB.2 Construction of Out-Of-Sample Data

The construction of the Out-Of-Sample Data on which the Machine Learning from the Training Data is applied towards, is similar to the construction of the Training Data in that we first identify the individuals affiliated with a set of universities, gather their work histories and generate the 1,400+ characteristics we tested against. In this case, the documentation will focus on the set of universities that we identify.

Our frame begins with Carnegie Institute Ph.D. granting research institutions, of which there exists approximately 230 institutions. These are all 4-year universities that have graduate and Ph.D. programs, with some possessing medical schools, some public, some private, etc… The important thing is that the institution contains undergraduates, graduate students and post-docs, along with faculty. We exclude liberal art colleges, medical schools only (no undergraduate) and other non-Ph.D. granting institutions. From the list of 230, we then rank them by R&D funds gathered from the NSF. We rank them according to the 2014 total R&D allocation and sum up

4 NSF data is compiled from https://ncsesdata.nsf.gov/profiles/site?method=rankingBySource&ds=herd

the totals across all universities. We then only keep the top 90% of institutions by research funds, meaning that we only keep the largest R&D universities that contribute up to 90% of total University R&D outlays. The remaining universities are dropped. This gives us a set of 130 research universities. The cutoff R&D expenditure value was approximately $100,000,000. Out of the $65 Billion spent by Universities in 2014, our set of universities spent approximately $52.5Billion (equivalent to 80%) and our universities contribute nearly 90% of NIH spending and provide wide coverage. Below are some summary statistics of the universities in our sample:

University Summary Statistics, 130 Universities.

	130 Universities	UMETRIC Sample[5]
Mean R&D Expenditure ($000), 2014	424,600	661,700
Mean Non-R&D Expenditures ($000), 2014	20,400	35,800
Mean # of NIH Awards, 2014	270	440
Mean Annual Enrollment	30,800	43,400
Mean Amount of NIH Awards ($000), 2014	112,500	180,900
Mean Undergraduate Enrollment, 2014	19,800	27,700
Mean Bachelor Degrees Awarded, 2014	4,700	6,900
Mean Graduate Enrollment, 2014	7,900	11,800
Mean Master Degrees Awarded, 2014	1,900	3,100
Mean Doctoral Degrees Awarded, 2014	700	1,100
Mean Total Degrees Awarded, 2014	7,300	11,100
Mean Faculty Number, 2014	1,400	2,200
% Private	28.5	30.8
% Land Grant	40	61.5
% with Medical School	69.2	84.6
Mean SAT Combined, 2014	1,140	1,190

As we can see, the majority of universities included in the sample are large, public universities with medical schools attached to them. The UMETRIC sample is slightly larger on average and expends more on R&D.

Once we apply these universities to the construction, we end up with approximately 7.3M Out-of-Sample individuals affiliated with universities between 2005 and 2014.

Given the approximate 10% UMETRIC rate for the In-Sample Training data (143,000 out of 1.5M), and given that average school in the UMETRIC sample spends more on R&D, we expect that the total Out-of-Sample Imputed UMETRIC size to be between 400,000 to 600,000.

[5] Excludes SUNY and University of Hawaii and University of Kansas, which did not make the list of Top 130 Research Universities

Appendix II: Full Regression Results

The tables below give the full regression results for the specification listed earlier. Tables 9, 10 and 11 include regressions by Size category.

Table 8A Full Results – OLS on All Startup Outcomes

Outcome Variable	Survival, t+1	Success, t+5	High-Growth (Employment) Success, t+5	High-Growth (Revenue) Success, t+5	High-Growth (Revenue per Employee) Success, t+5
ln $EARN_{f0}$	0.0764***	0.0267***	0.00541***	0.00472***	0.00337***
	(0.000176)	(0.000215)	(0.000101)	(0.000129)	(0.000127)
ln $\overline{AGE}_{f0}$	-0.00389***	-0.0187***	0.000352	-0.0119***	-0.00211***
	(0.000658)	(0.000782)	(0.000367)	(0.000469)	(0.000462)
$SIZE\ DUMMIES$	Yes	Yes	Yes	Yes	Yes
ln $FEMALE_{f0}$	0.0117***	0.0370***	0.00763***	-0.00184***	-0.00376***
	(0.000238)	(0.000278)	(0.000131)	(0.000167)	(0.000165)
ln $FOREIGN_{f0}$	0.00939***	0.0179***	0.00568***	0.00202***	0.000792***
	(0.000256)	(0.000300)	(0.000141)	(0.000180)	(0.000177)
ln RD_{f0}	-0.00200***	0.0260***	0.0107***	0.00256***	-0.00258***
	(0.000372)	(0.000472)	(0.000222)	(0.000283)	(0.000279)
ln $RD \times FEMALE_{f0}$	0.00680***	-0.0143***	-0.00329***	-0.000359	0.000139
	(0.000469)	(0.000601)	(0.000282)	(0.000360)	(0.000355)
ln $RD \times FOREIGN_{f0}$	-0.00307***	-0.0124***	0.000645	-0.00220***	-0.000283
	(0.000578)	(0.000749)	(0.000352)	(0.000449)	(0.000443)
ln HT_{f0}	-0.00893***	0.0198***	0.00790***	0.00600***	-0.00117**
	(0.000486)	(0.000616)	(0.000289)	(0.000369)	(0.000364)
ln $HT \times FEMALE_{f0}$	0.00232***	-0.0142***	-0.00105**	-0.00215***	0.00116*
	(0.000658)	(0.000847)	(0.000398)	(0.000508)	(0.000501)
ln $HT \times FOREIGN_{f0}$	-0.00823***	-0.00277**	0.00272***	0.00152**	-0.000597
	(0.000710)	(0.000884)	(0.000415)	(0.000530)	(0.000523)
ln UNI_{f0}	0.00433***	0.0220***	0.00866***	0.00274***	0.000258
	(0.000803)	(0.00111)	(0.000521)	(0.000666)	(0.000656)
ln $UNI \times FEMALE_{f0}$	0.00829***	0.00420**	0.000811	0.000441	-0.00103
	(0.000988)	(0.00137)	(0.000642)	(0.000819)	(0.000808)
ln $UNI \times FOREIGN_{f0}$	-0.00837***	-0.00543**	0.00451***	0.00273*	0.00245*
	(0.00133)	(0.00183)	(0.000860)	(0.00110)	(0.00108)
ln $UMETRIC_{f0}$	0.0143***	0.00578	0.00809***	0.00306	0.000259
	(0.00299)	(0.00367)	(0.00172)	(0.00220)	(0.00217)
ln $UMETRIC \times FEMALE_{f0}$	-0.00223	-0.00353	-0.00996***	-0.00370	0.0000737
	(0.00381)	(0.00464)	(0.00218)	(0.00278)	(0.00274)
ln $UMETRIC \times FOREIGN_{f0}$	-0.0189***	0.000354	0.00700*	0.00605	-0.00195
	(0.00524)	(0.00663)	(0.00311)	(0.00397)	(0.00392)
Zip Code-Year FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Constant	0.284***	-0.297***	-0.0963***	-0.000259	-0.00160
	(0.0380)	(0.0733)	(0.0281)	(0.0359)	(0.0354)
Observations	3,730,000	3,730,000	3,730,000	3,730,000	3,730,000
R-squared	0.079	0.193	0.049	0.018	0.014

Robust Standard Errors in Parentheses. *p<0.05, **p<0.01, ***p<0.001;

Table 9A Full Results – OLS on All Startup Growth Rates

Outcome Variable	Employment Growth, t+1	Revenue Growth, t+1	Revenue per Employee Growth, t+1
$\ln EARN_{f0}$	0.154***	0.0260***	0.0316***
	(0.000879)	(0.000700)	(0.000805)
$\ln \overline{AGE}_{f0}$	-0.0969***	-0.0167***	0.0370***
	(0.00338)	(0.00243)	(0.00277)
$SIZE\ DUMMIES$	Yes	Yes	Yes
$\ln FEMALE_{f0}$	0.234***	0.00375***	-0.0869***
	(0.000937)	(0.000854)	(0.000971)
$\ln FOREIGN_{f0}$	0.0606***	0.00951***	-0.0292***
	(0.000786)	(0.000902)	(0.00102)
$\ln RD_{f0}$	0.151***	-0.0147***	-0.107***
	(0.00117)	(0.00139)	(0.00157)
$\ln RD \times FEMALE_{f0}$	-0.102***	-0.00694***	0.0305***
	(0.00134)	(0.00175)	(0.00198)
$\ln RD \times FOREIGN_{f0}$	-0.0419***	0.00759***	0.0319***
	(0.00151)	(0.00214)	(0.00242)
$\ln HT_{f0}$	0.0483***	0.00636***	-0.0501***
	(0.00145)	(0.00177)	(0.00201)
$\ln HT \times FEMALE_{f0}$	-0.0352***	0.00189	0.0292***
	(0.00183)	(0.00242)	(0.00274)
$\ln HT \times FOREIGN_{f0}$	-0.0121***	-0.00432	-0.000118
	(0.00195)	(0.00251)	(0.00284)
$\ln UNI_{f0}$	0.0517***	0.0232***	-0.0347***
	(0.00219)	(0.00307)	(0.00348)
$\ln UNI \times FEMALE_{f0}$	-0.0336***	-0.00339	0.0175***
	(0.00260)	(0.00376)	(0.00425)
$\ln UNI \times FOREIGN_{f0}$	-0.0247***	-0.00550	0.00991
	(0.00341)	(0.00491)	(0.00555)
$\ln UMETRIC_{f0}$	0.0177*	0.00109	-0.0131
	(0.00739)	(0.0110)	(0.0124)
$\ln UMETRIC \times FEMALE_{f0}$	-0.0162	-0.00328	0.0128
	(0.00926)	(0.0139)	(0.0157)
$\ln UMETRIC \times FOREIGN_{f0}$	0.0182	0.0272	0.0231
	(0.0133)	(0.0192)	(0.0216)
Zip Code-Year FE	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes
Constant	-2.771***	-0.536	-0.914
	(0.0149)	(0.485)	(9332.2)
Observations	723105	1472946	1436529
R-squared	0.201	0.027	0.057

Robust Standard Errors in Parentheses. *p<0.05, **p<0.01, ***p<0.001;

Table 10A: Full Results of OLS on Industrial Startup Outcomes, 2005-2014

Outcome Variable	Survival, t+1	Success, t+5	High-Growth (Employment) Success, t+5	High-Growth (Revenue) Success, t+5	High-Growth (Revenue per Employee) Success, t+5
Sample	Industrial Startups	Industrial Startups	Industrial Startups	Industrial Startups	Industrial Startups
$\ln EARN_{f0}$	0.0546***	0.0229***	0.00387***	0.00338***	0.00206***
	(0.000265)	(0.000381)	(0.000195)	(0.000262)	(0.000231)
$\ln \overline{AGE}_{f0}$	0.00233*	-0.0137***	0.00552***	-0.0183***	-0.00391***
	(0.00110)	(0.00154)	(0.000791)	(0.00106)	(0.000936)
$SIZE\ DUMMIES$	Yes	Yes	Yes	Yes	Yes
$\ln FEMALE_{f0}$	0.00623***	0.0524***	0.00914***	-0.00397***	-0.00796***
	(0.000405)	(0.000557)	(0.000286)	(0.000383)	(0.000338)
$\ln FOREIGN_{f0}$	0.00310***	0.0172***	0.00837***	0.00444***	0.000843*
	(0.000476)	(0.000654)	(0.000336)	(0.000450)	(0.000397)
$\ln RD_{f0}$	-0.0183***	0.0273***	0.0151***	0.00544***	-0.00218***
	(0.000700)	(0.00105)	(0.000541)	(0.000725)	(0.000640)
$\ln RD \times FEMALE_{f0}$	0.0162***	-0.0148***	-0.00113	-0.00146	0.000143
	(0.000838)	(0.00128)	(0.000654)	(0.000878)	(0.000775)
$\ln RD \times FOREIGN_{f0}$	0.000894	-0.0186***	0.00547***	-0.00188	0.00104
	(0.000984)	(0.00151)	(0.000776)	(0.00104)	(0.000919)
$\ln HT_{f0}$	-0.0128***	0.0343***	0.00914***	0.00863***	-0.00336***
	(0.000700)	(0.00100)	(0.000515)	(0.000691)	(0.000610)
$\ln HT \times FEMALE_{f0}$	0.00720***	-0.0387***	-0.00371***	-0.00281**	0.00570***
	(0.000947)	(0.00139)	(0.000711)	(0.000953)	(0.000841)
$\ln HT \times FOREIGN_{f0}$	0.000230	-0.000839	-0.000794	-0.000658	-0.00123
	(0.000926)	(0.00131)	(0.000670)	(0.000899)	(0.000793)
$\ln UNI_{f0}$	-0.00408***	0.0289***	0.0123***	0.00730***	0.000931
	(0.00124)	(0.00208)	(0.00107)	(0.00143)	(0.00127)
$\ln UNI \times FEMALE_{f0}$	0.0125***	0.000674	0.00190	-0.00227	-0.000243
	(0.00150)	(0.00252)	(0.00129)	(0.00173)	(0.00153)
$\ln UNI \times FOREIGN_{f0}$	-0.00178	0.000760	0.00548***	0.00144	0.00211
	(0.00183)	(0.00304)	(0.00156)	(0.00209)	(0.00184)
$\ln UMETRIC_{f0}$	0.00604	0.00578	0.0177***	0.0127*	0.00596
	(0.00474)	(0.00738)	(0.00379)	(0.00508)	(0.00448)
$\ln UMETRIC \times FEMALE_{f0}$	0.000158	0.0121	-0.0141**	-0.00996	-0.00654
	(0.00598)	(0.00918)	(0.00471)	(0.00631)	(0.00557)
$\ln UMETRIC \times FOREIGN_{f0}$	-0.00394	-0.00706	-0.00460	0.00307	0.00175
	(0.00692)	(0.0110)	(0.00564)	(0.00757)	(0.00668)
Zip Code-Year FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Constant	0.532***	0.498***	-0.141*	0.0284	0.0231
	(0.116)	(0.117)	(0.0602)	(0.0807)	(0.0713)
Observations	1013341	517087	516902	516928	516928
R-squared	0.069	0.213	0.079	0.024	0.019

Robust Standard Errors in Parentheses. *p<0.05, **p<0.01, ***p<0.001

Table 11A: Full Results of OLS on High-Tech Startup Outcomes, 2005-2014

Outcome Variable	Survival, t+1	Success, t+5	High-Growth (Employment) Success, t+5	High-Growth (Revenue) Success, t+5	High-Growth (Revenue per Employee) Success, t+5
Sample	High-Tech Startups	High-Tech Startups	High-Tech Startups	High-Tech Startups	High-Tech Startups
$\ln EARN_{f0}$	0.0555***	0.0245***	0.00617***	0.00592***	0.000307
	(0.000784)	(0.00109)	(0.000573)	(0.000926)	(0.000745)
$\ln \overline{AGE}_{f0}$	-0.00637*	-0.0280***	-0.00166	-0.0295***	-0.00274
	(0.00319)	(0.00435)	(0.00229)	(0.00369)	(0.00297)
$SIZE\ DUMMIES$	Yes	Yes	Yes	Yes	Yes
$\ln FEMALE_{f0}$	-0.00799	0.0890***	0.0179***	0.00887	-0.0105*
	(0.00445)	(0.00634)	(0.00333)	(0.00539)	(0.00434)
$\ln FOREIGN_{f0}$	0.0143*	0.0667***	0.00817	0.0152*	0.00800
	(0.00597)	(0.00857)	(0.00451)	(0.00728)	(0.00586)
$\ln RD_{f0}$	-0.00704***	0.0308***	0.0173***	0.0103***	-0.00333*
	(0.00165)	(0.00241)	(0.00127)	(0.00205)	(0.00165)
$\ln RD \times FEMALE_{f0}$	-0.00749**	-0.0167***	-0.000627	-0.00233	-0.000377
	(0.00238)	(0.00354)	(0.00186)	(0.00301)	(0.00242)
$\ln RD \times FOREIGN_{f0}$	-0.00687**	-0.0171***	0.00443*	-0.00477	0.00326
	(0.00243)	(0.00357)	(0.00188)	(0.00304)	(0.00244)
$\ln HT_{f0}$	-0.0342***	0.0679***	0.0114***	0.00949***	-0.00674***
	(0.00165)	(0.00221)	(0.00116)	(0.00188)	(0.00151)
$\ln HT \times FEMALE_{f0}$	0.0281***	-0.0889***	-0.0180***	-0.0148**	0.0101*
	(0.00465)	(0.00659)	(0.00347)	(0.00560)	(0.00451)
$\ln HT \times FOREIGN_{f0}$	-0.00558	-0.0483***	-0.00138	-0.00948	-0.00773
	(0.00604)	(0.00865)	(0.00455)	(0.00735)	(0.00592)
$\ln UNI_{f0}$	0.00627*	0.00959*	0.00440	0.000488	0.00453
	(0.00271)	(0.00438)	(0.00230)	(0.00372)	(0.00299)
$\ln UNI \times FEMALE_{f0}$	0.00246	-0.0133*	-0.000512	0.00161	-0.000307
	(0.00398)	(0.00640)	(0.00337)	(0.00544)	(0.00438)
$\ln UNI \times FOREIGN_{f0}$	0.00331	0.0269***	0.0237***	0.00445	-0.00715
	(0.00416)	(0.00648)	(0.00340)	(0.00550)	(0.00443)
$\ln UMETRICS_{f0}$	0.00984	0.0185	0.0201**	0.0313*	0.0127
	(0.00968)	(0.0145)	(0.00764)	(0.0124)	(0.00994)
$\ln UMETRICS \times FEMALE_{f0}$	0.0275	0.00251	-0.0268*	-0.0358	-0.00263
	(0.0157)	(0.0237)	(0.0124)	(0.0201)	(0.0162)
$\ln UMETRICS \times FOREIGN_{f0}$	-0.0233	-0.0362	-0.0144	-0.0218	-0.00148
	(0.0143)	(0.0215)	(0.0113)	(0.0183)	(0.0147)
Zip Code-Year FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Constant	0.498*	0.471	-0.0864	0.0276	0.0418
	(0.251)	(0.257)	(0.134)	(0.218)	(0.176)
Observations	129256	69521	69505	69510	69510
R-squared	0.112	0.193	0.133	0.067	0.057

Robust Standard Errors in Parentheses. *p<0.05, **p<0.01, ***p<0.001