Web Appendix

A. Data Appendix

In this appendix, we describe the construction of the firm-level revenue variable that serves as the basis for our analysis. We then describe how this variable is used to construct a revenue enhanced subset of the LBD that includes continuers, births and deaths and discuss our methodology for cleaning the data. Finally, we describe our implementation of propensity score matching to control for potential selection effects. In presenting our propensity score models, we compare propensity score adjusted job creation and job destruction statistics from the revenue enhanced subset to the results for the full LBD to indicate the effectiveness of our strategy.

I. Construction of the Revenue Variable

The U.S. Census Bureau Business Register files contain revenue data sourced from business administrative income and payroll filings. These data are used for statistical purposes including the Economic Census program and the Nonemployer Statistics program. There are a number of different tax forms and different revenue items within those forms that are relevant for calculating firm-level revenue depending on the sector that a firm operates in (or more specifically, the particular reporting tax unit, the EIN, within a firm), as well as the legal form of organization of the firm (nonprofits, partnerships, corporations, or sole proprietors). In an effort to build revenue measures reasonably comparable across firms, starting in 2002 the Census developed an algorithm that takes these differences in tax forms and revenue concepts into account.[1] Within the Census, this "best receipts" variable has previously been applied to single-

[1] Algorithms are available to Census Bureau employees and RDC researchers that have an approved project and a need to know. Depending on the form and industry these algorithms may include total revenue, net receipts, gross revenue, receipts from interest, receipts from gross rents, total income, cost of goods sold, and direct as well as rent expenses.

unit firms only. Thus, we extended the original methodology in two ways. First, we apply the Census Bureau methodology to multi-unit firms. Multi-establishment firms can report different parts of their operations under different and independent EIN filings. There are many possible reasons may organize across multiple EINs including geographic, tax status, or business considerations. Given these within-firm sources of variation, we apply the algorithm at the EIN level first, using the EIN's self reported NAICS classification to assign an industry to the EIN. The taxable revenue items that are included in the EINs total revenue are determined by this industry designation. We then compute a firm-level revenue measure by summing up all of the EINs associated with a particular firm.

Second, we developed an analog of the algorithm for years prior to 2002. The Business Register went through a complete redesign in 2002 which made it possible to keep additional fields that had been combined in prior years. We modify the pre-2000 algorithm to adjust for the different revenue items available before 2002. For any given year of revenue, we use prior year revenue variables from the following year's BR. Previous research from the Census has indicated that due to extended filing schedules, late filing, and other factors, these prior year revenue variables provide significantly improved revenue information. Thus, in applying our algorithms we always use revenue for a given year from the BR file for the following year. Figure A1 shows the results of applying these algorithms on the BR revenue measures and after filtering. Revenue is deflated using the GDP Implicit Price Deflator. Real revenue is in 2009 dollars.

II. The Revenue Enhanced LBD Subset

Based on the revenue variable describe above, each observation in the LBD falls into one of four revenue categories: revenue continuers with revenue data in both year *t-1* and year *t*,

revenue deaths with revenue in year *t-1* but no revenue in year *t*, revenue births with no revenue in year *t-1* and revenue in year *t*, and observations with no revenue data at either time. Observations in the fourth category are dropped in their entirety from the sample, while the subsets represented by the first three categories are cleaned to insure that the observations are suitable for analysis.

Inspection of the revenue data reveals a number of outliers. These can come about for a number of reasons including typographical errors, OCR errors, units errors, and even denomination errors. Outliers are also common amongst commodity and energy trading entities as well as businesses organized in terms of holding companies. To address these issues, for the revenue continuers subset we apply the following filters:

(1) We drop observations with labor productivity (revenue divided by employment) above the 99.9th percentile and below the 0.1th percentile for both years *t-1* and *t*.

(2) We drop observations reporting over $1 billion in average revenue and a DHS revenue growth rate of less then -0.5 or greater than 0.5.

(3) We drop observations reporting over $100 million in average revenue and a DHS revenue growth rate of less than -1.5 or greater than 1.5.

(4) We drop any observations reporting $1 trillion in average revenue or more.

These filters are designed to narrowly target specific problems such as unusually high or low labor productivity values, unusually high revenue values, and unusually high changes in revenue all the while minimizing the number of records we exclude from the data. Overall, this procedure excludes 0.14% of the total universe of revenue continuers.

For the revenue deaths and births we apply the same labor productivity filter for the relevant year of revenue. Because all revenue deaths and births have DHS revenue growth rates of -2 or

2, application of the additional filters amounts to a restriction on the DHS revenue denominator of $100 million. Overall, this procedure excludes 0.08% and 0.13% of the total universe of revenue deaths and births respectively. Then, so that only true employment deaths and births are counted, the revenue death and revenue birth files are restricted to observations that represent employment deaths for the former and employment births for the latter. The remaining observations from each subset are then combined to form the revenue enhanced LBD subset.

III. Missing Observations, Selection, and Propensity Score Matching

Firms typically use the same EINs when filing payroll and income tax reports. This facilitates linking employment and revenue activity for a given firm at the Census Bureau. However, this is not always the case. About 20 percent of businesses file their payroll and income reports under different EINs. When this happens, the Census Bureau has no direct way of linking the two records. These revenue EINs become orphan records to payroll EINs although they are never identified as such. Revenue records without a corresponding payroll record are considered non-employer EINs.[2] The practical consequence of this is that for 21.8 percent of the revenue enhanced LBD subset, we are missing revenue data. Further, it is often the case that employers will consistently use different EINs when filing their payroll and income so many of these firms are missing all of their revenue data making it difficult to impute their records. In addition to potential selection resulting from the examination of only observations that have revenue data, the additional filters and restrictions placed on the data may create problematic selection effects, particularly in the case of deaths and births.

Given that selection effects may differ for continuers, deaths, and births, we developing separate propensity score models for employment continuers with revenue data, employment

[2] For example corporations file form 1120 for their income taxes and form 941 for their employment taxes. http://www.irs.gov/Businesses/Small-Businesses-&-Self-Employed/Corporations

deaths with revenue data, and employment births with revenue data. Each of these partitions constitutes the set of firms for which the dependent variable equals one in a propensity score model that is run on the universe of LBD employment continuers, LBD employment deaths, and LBD employment births respectively. For the employment continuers, the propensity score is inverse probability weight calculated from the predicted values from a logistic regression including firm size, firm size squared, firm age, firm age squared, an indicator variable for firms of age 16+, employment growth rate (7 classes), broad industry (20 classes), and a multi-unit status indicator. For deaths, we employ the same model, except we exclude the growth rate classes. Finally, for births the model includes firm size, firm size squared, broad industry, and the indicator for multi-unit status. Figures A2-A4 examine the performance of our propensity score model in terms of total net job creation, job destruction from exit, and job creation from births. Although these figures indicate some modest selection effects present in the revenue enhanced LBD subset, the propensity score model yields patterns of employment growth dynamics for continuers, births and deaths for the enhanced revenue subset of the LBD that closely mimic those for the full LBD. Figures A2-A4 also show that even without weighting the enhanced revenue subset does a reasonable job of capturing the employment dynamics from the full LBD.

Figure A1 Real revenue (2009 dollars)

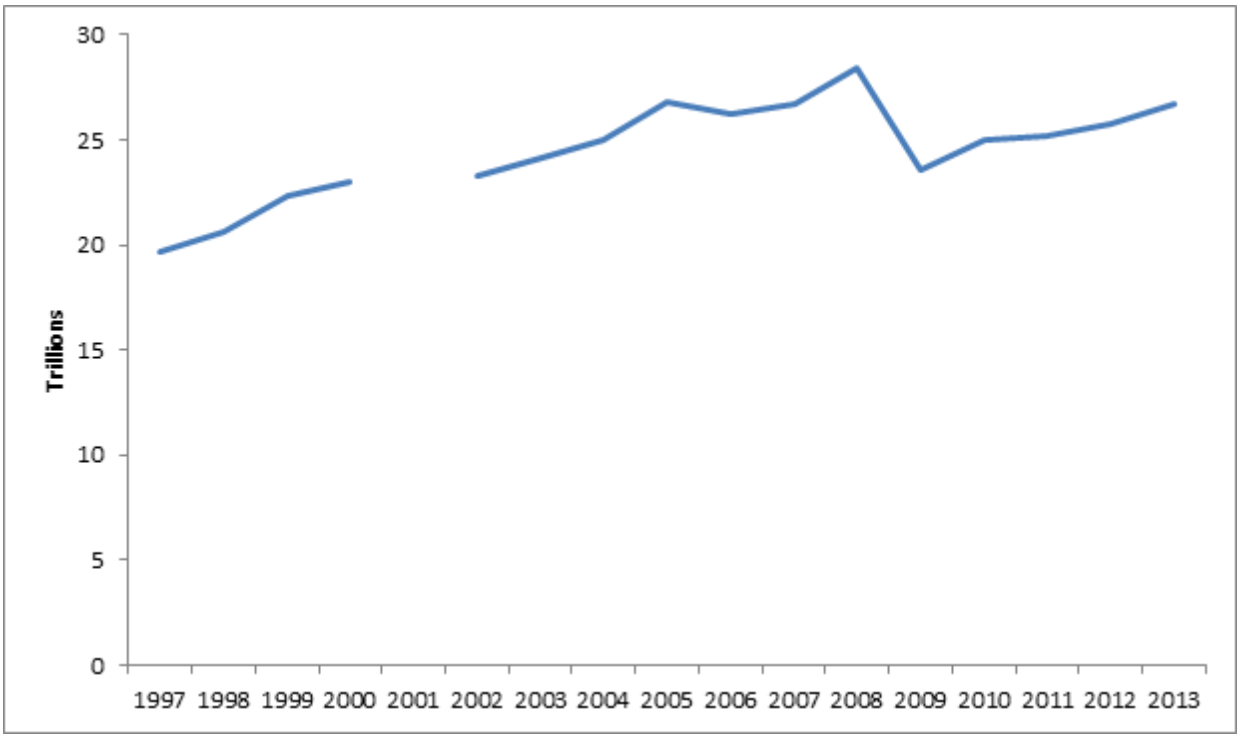


Figure A2 Net Employment Growth For Surviving Firms by Sample, 1996-2013

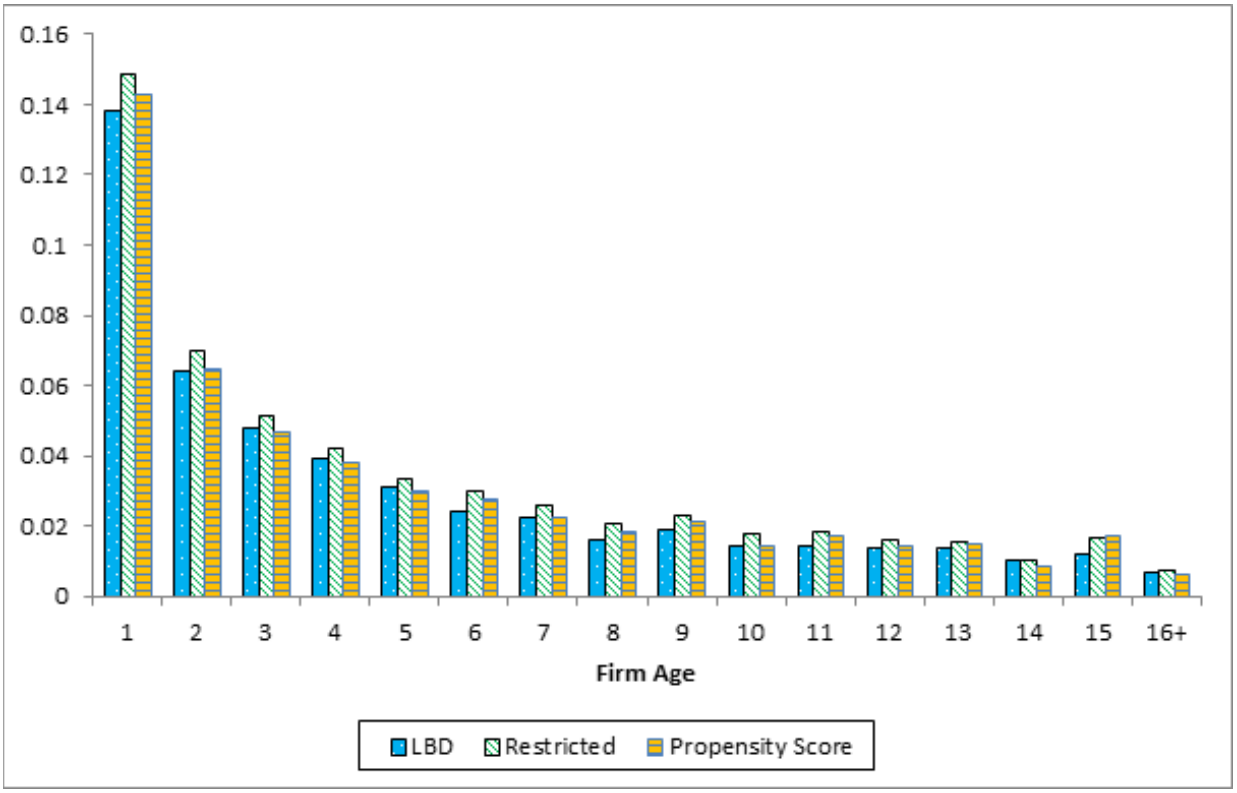


Figure A3 Job Destruction from Exit by Sample, 1996-2013

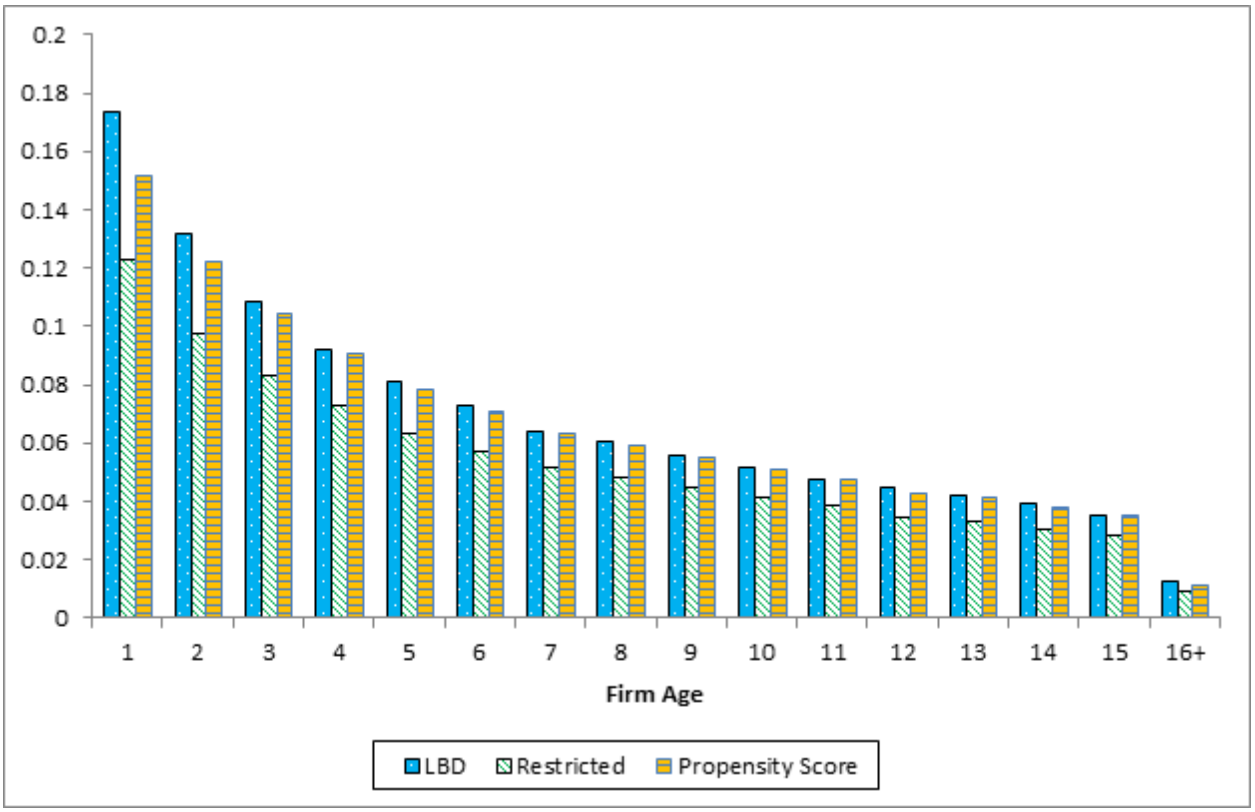
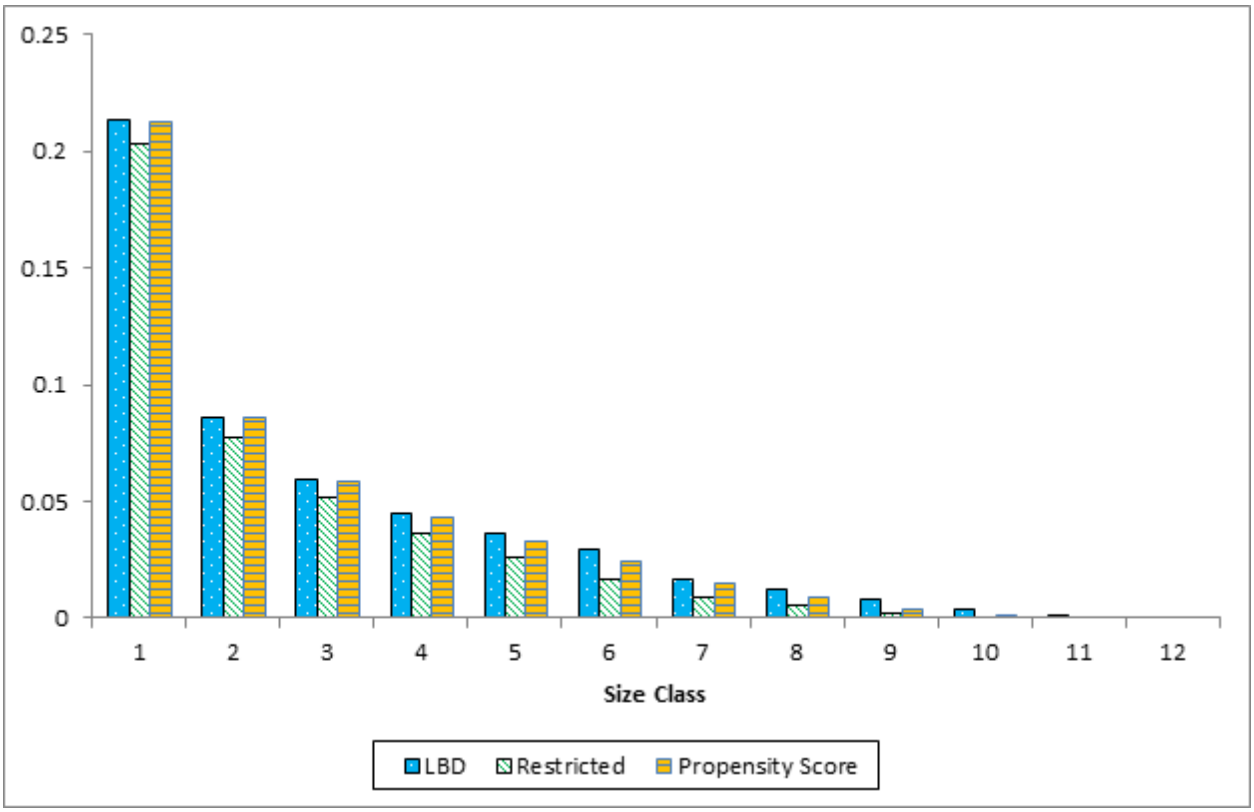
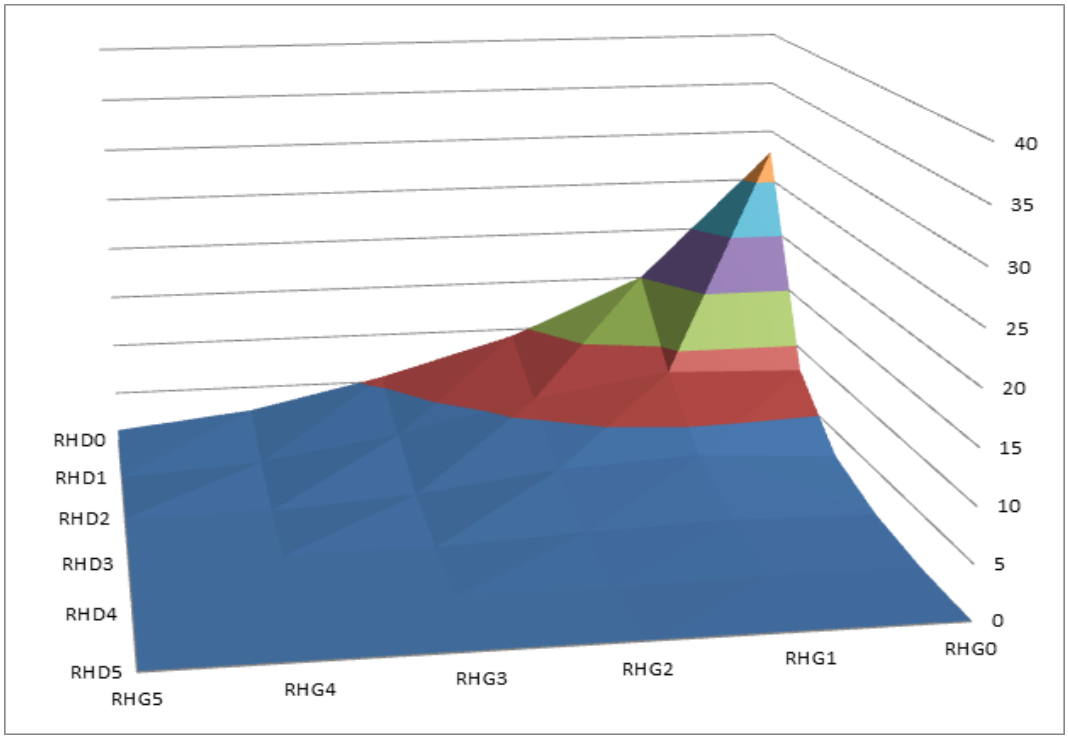


Figure A4 Job Creation from Births by Sample, 1996-2013



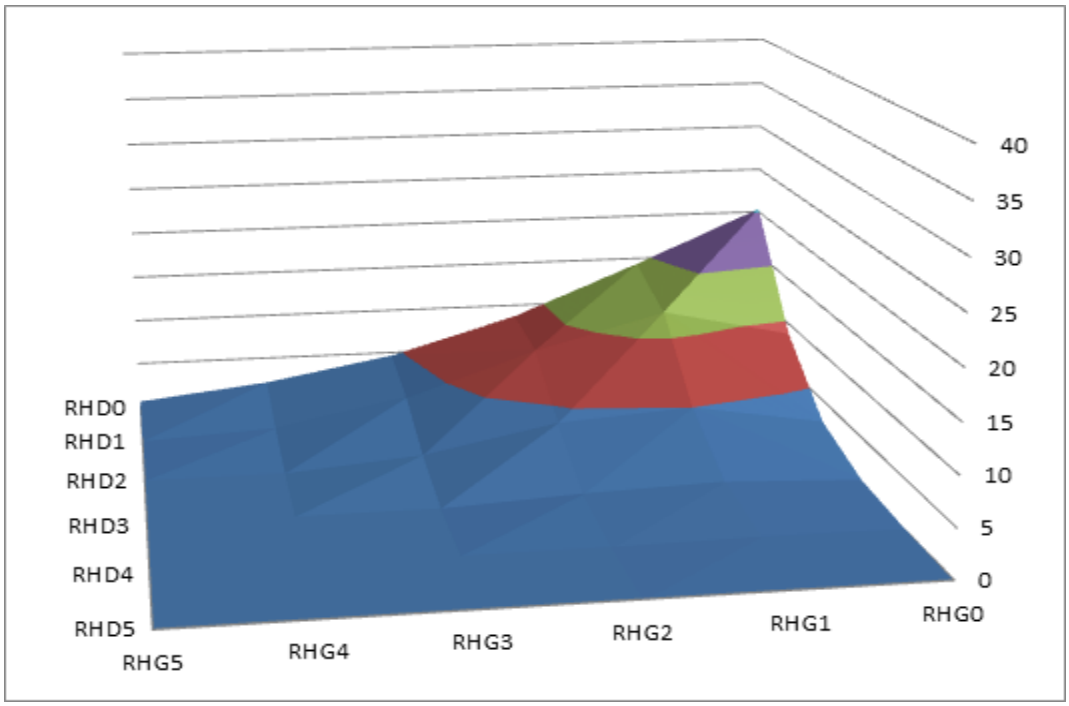
B. Supplemental Results

Figure B1 Percentage of RHG and RHD Events for 5 Year Old Firms (Revenue Weighted)



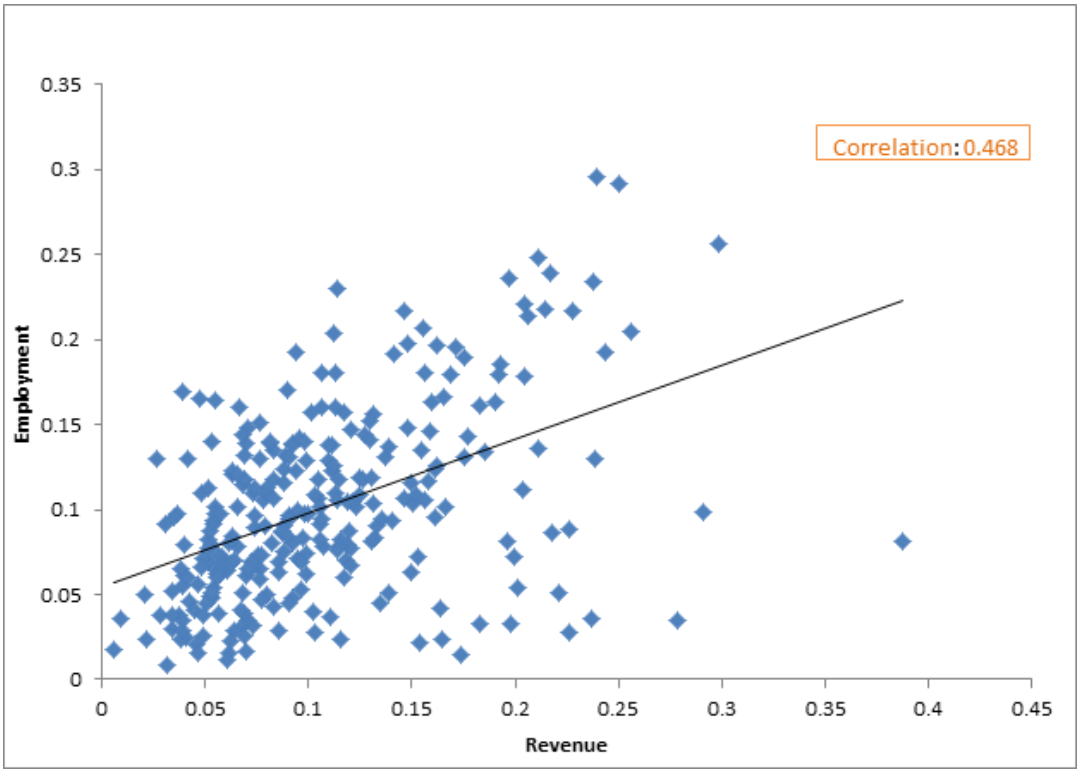
Source: Statistics computed from the Revenue enhanced LBD subset 1996-2000, 2003-2013. RHG = revenue high growth. RHD = revenue high decline. Reported shares are revenue weighted.

Figure B2 Percentage of EHG and EHD Events for 5 Year Old Firms (Employment Weighted)



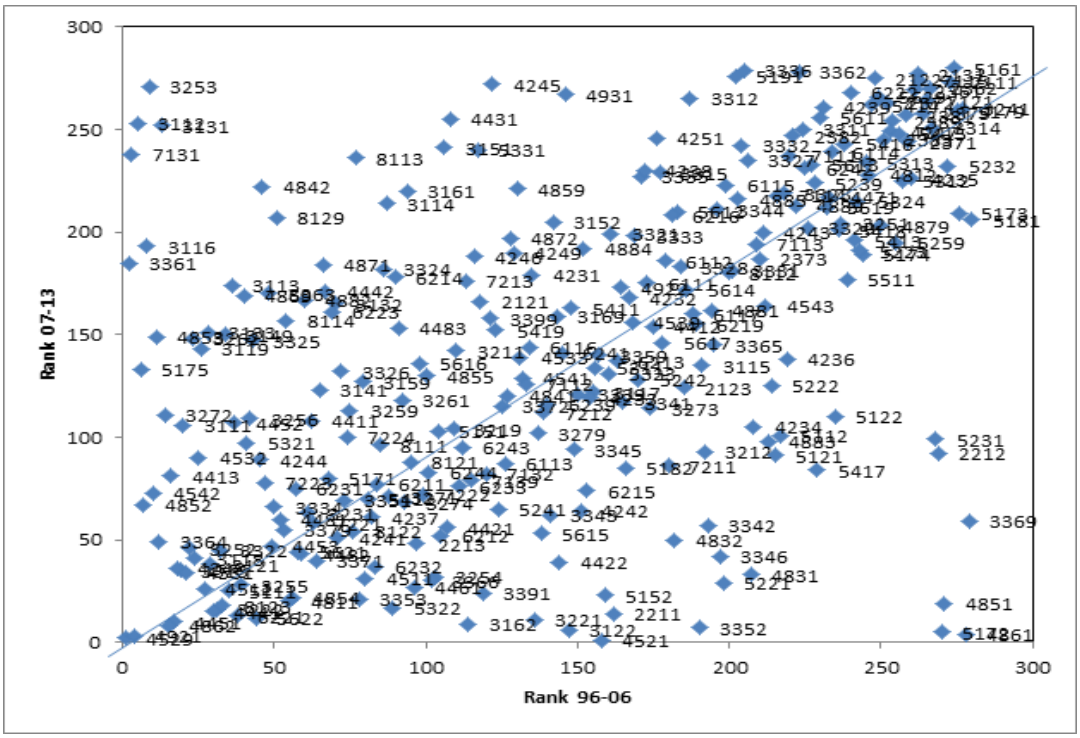
Source: Statistics computed from the Revenue enhanced LBD subset 1996-2000, 2003-2013. EHG = employment high growth. RHD = employment high decline. Reported shares are employment weighted.

Figure B3 Industry Effects Revenue versus Employment



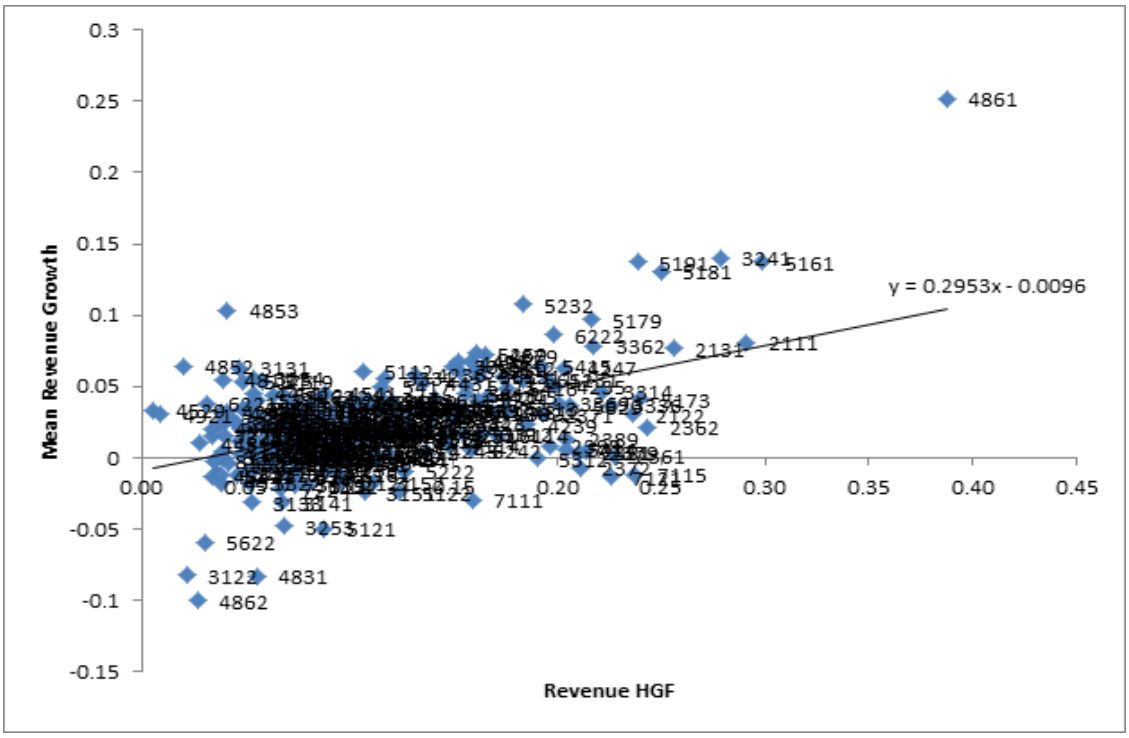
Source: Statistics computed from the Revenue enhanced LBD subset 1996-2000, 2003-2013. Reported are estimated effects of linear probability models on industry effects.

Figure B4 Industry Rankings in Revenue HGF, Change from 96-06 to 07-13



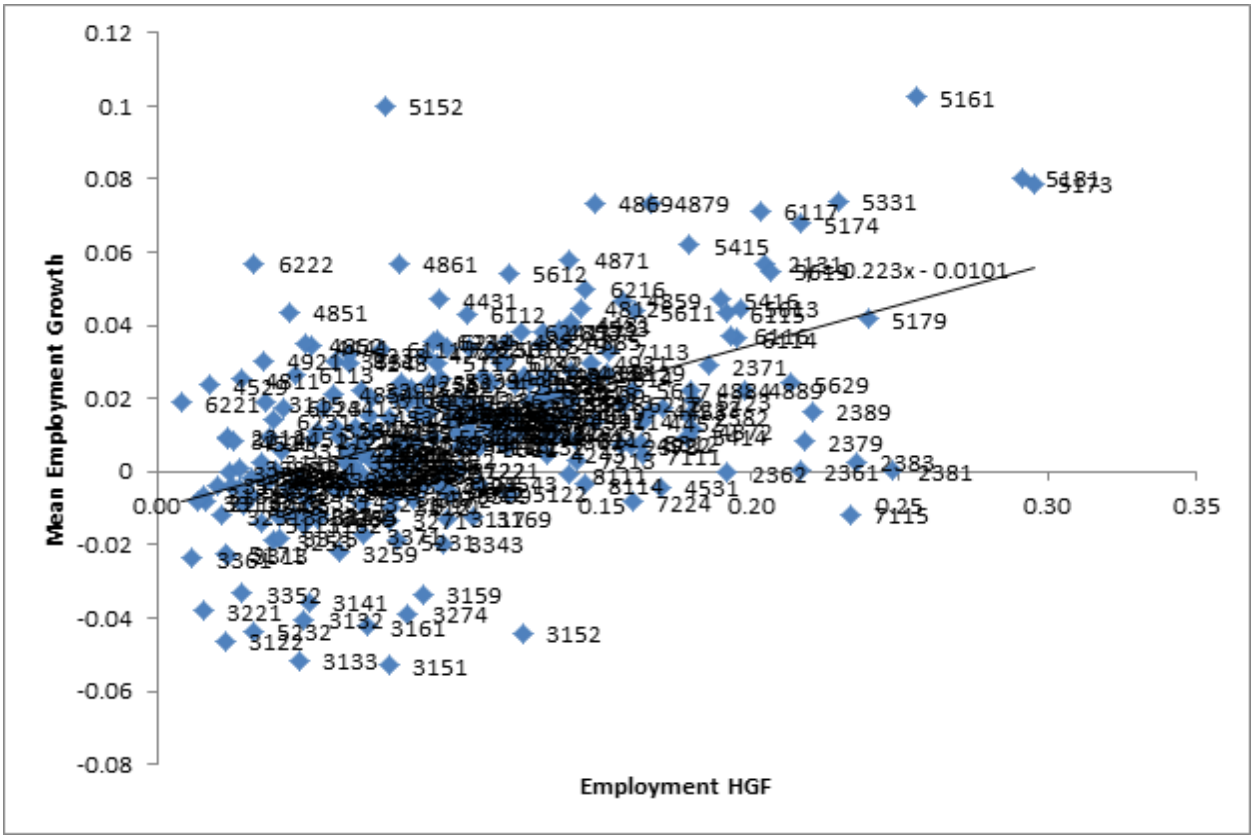
Source: Statistics computed from the Revenue enhanced LBD subset 1996-2000 and 2003-2013. The rankings for 1996-06 use the estimates from the 1996-06 period (except for 2001 and 2002) and the rankings of 2007-13 use the estimates from the 2007-13 period. Reported are estimated effects of linear probability models on industry effects.

Figure B5 Revenue HGF versus Revenue Growth by Industry



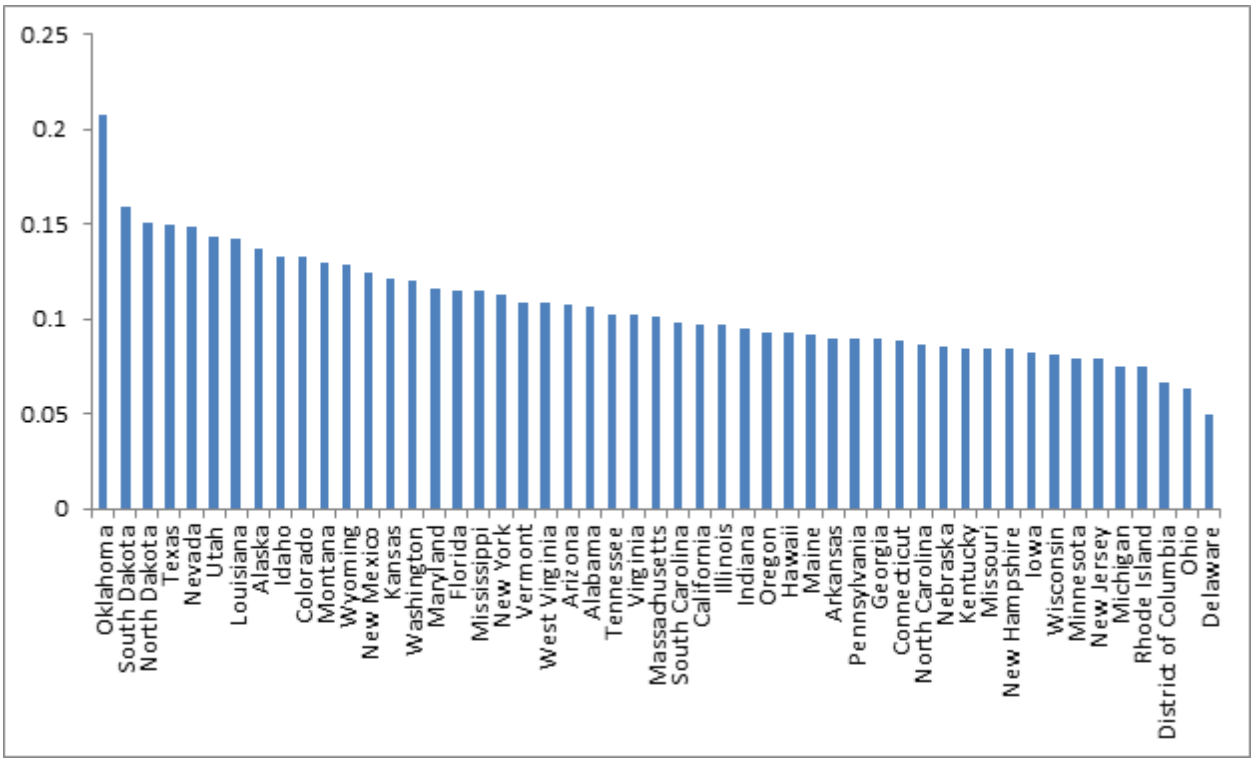
Source: Statistics computed from the Revenue enhanced LBD subset 1996-2000, 2003-2013. Reported Revenue HGF are estimated effects of linear probability models on industry effects. Mean Revenue Growth is revenue weighted mean revenue growth for firms.

Figure B6 Employment HGF versus Employment Growth by Industry



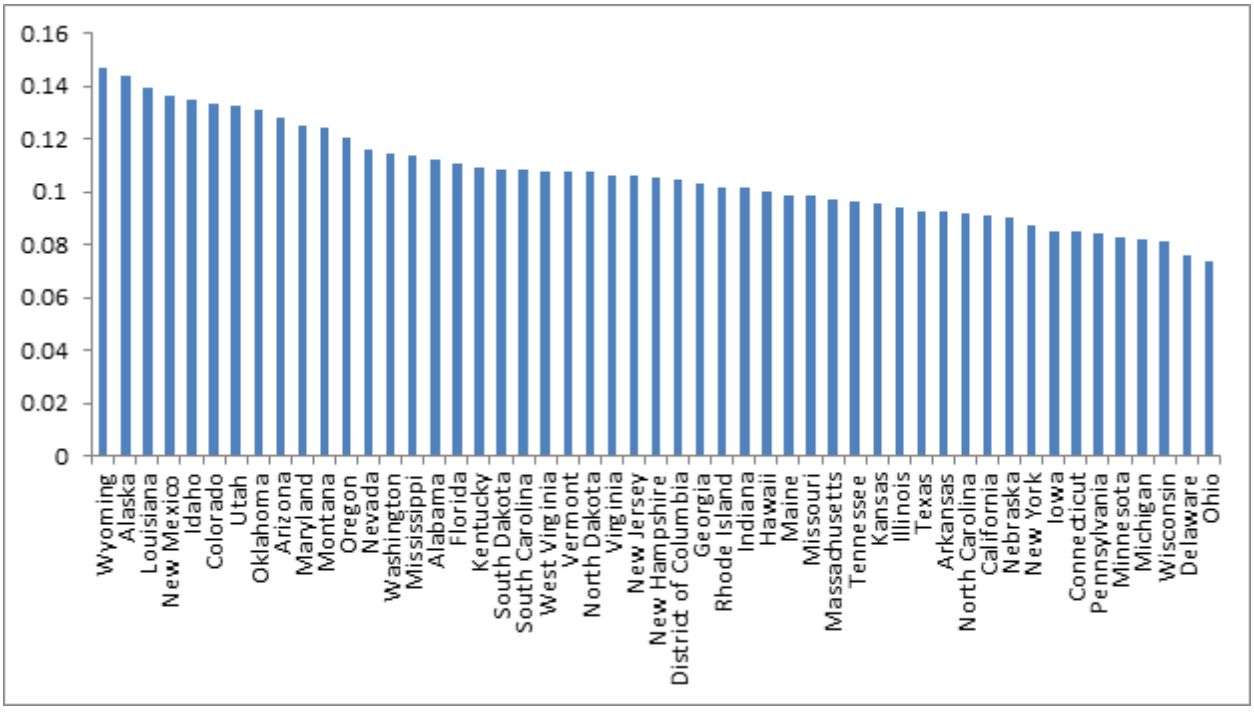
Source: Statistics computed from the Revenue enhanced LBD subset 1996-2000, 2003-2013. Reported Employment HGF are estimated effects of linear probability models on industry effects. Mean Employment growth is employment weighted mean employment growth for firms.

Figure B7 High Growth Firm State Effects (Revenue)



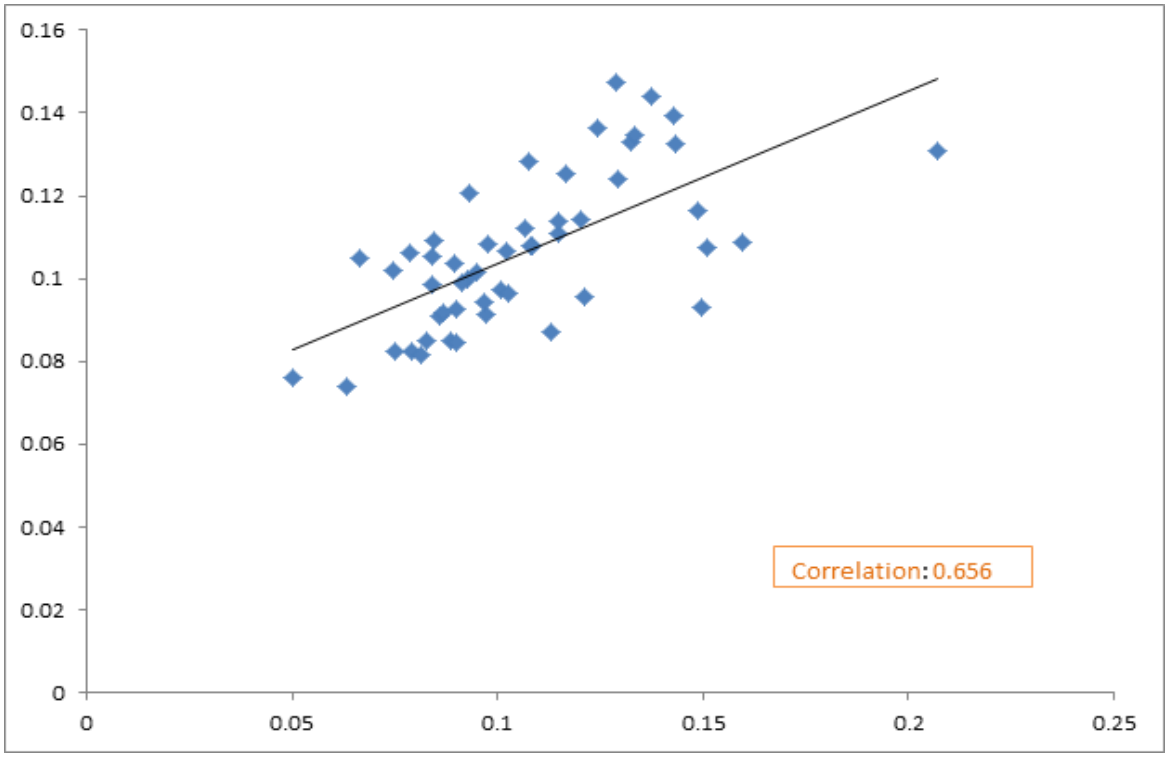
Source: Statistics computed from the Revenue enhanced LBD subset 1996-2000, 2003-2013. Reported are estimated effects of linear probability models on industry effects.

Figure B8 High Growth Firm State Effects (Employment)



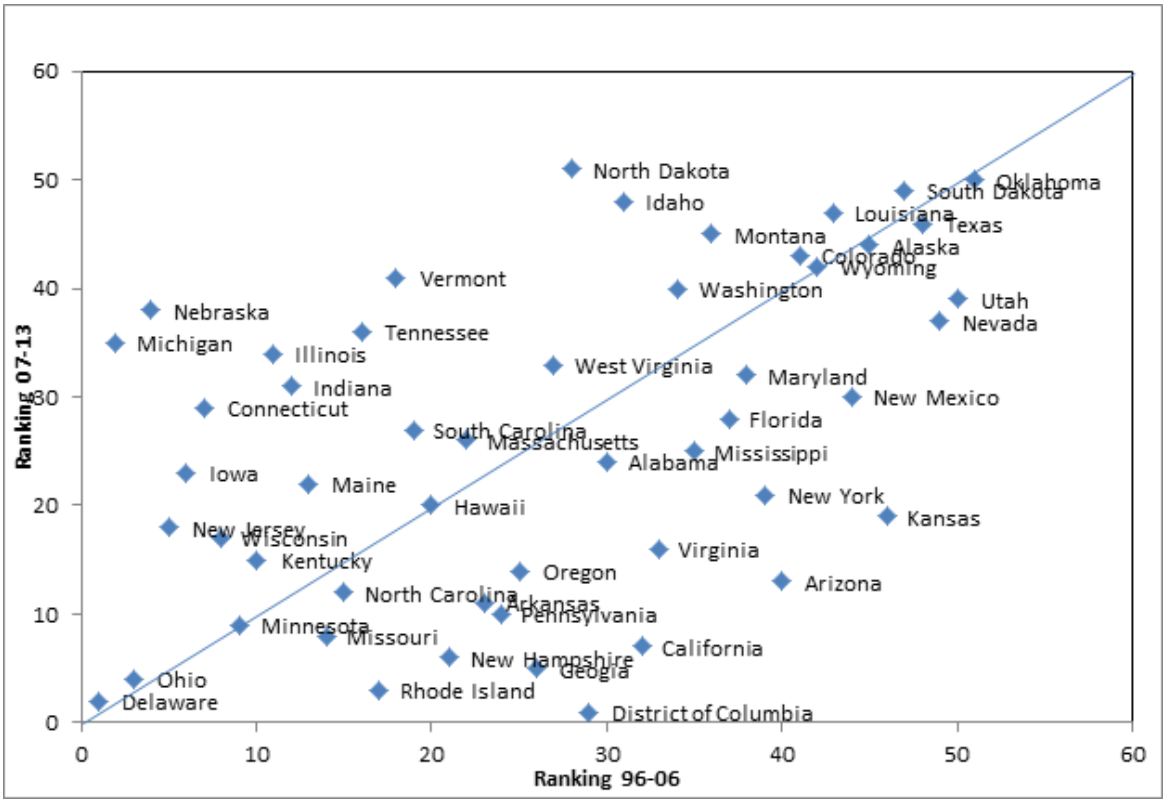
Source: Statistics computed from the Revenue enhanced LBD subset 1998-2000, 2003-2011. Reported are estimated effects of linear probability models on industry effects.

Figure B9 State Effects Revenue versus Employment



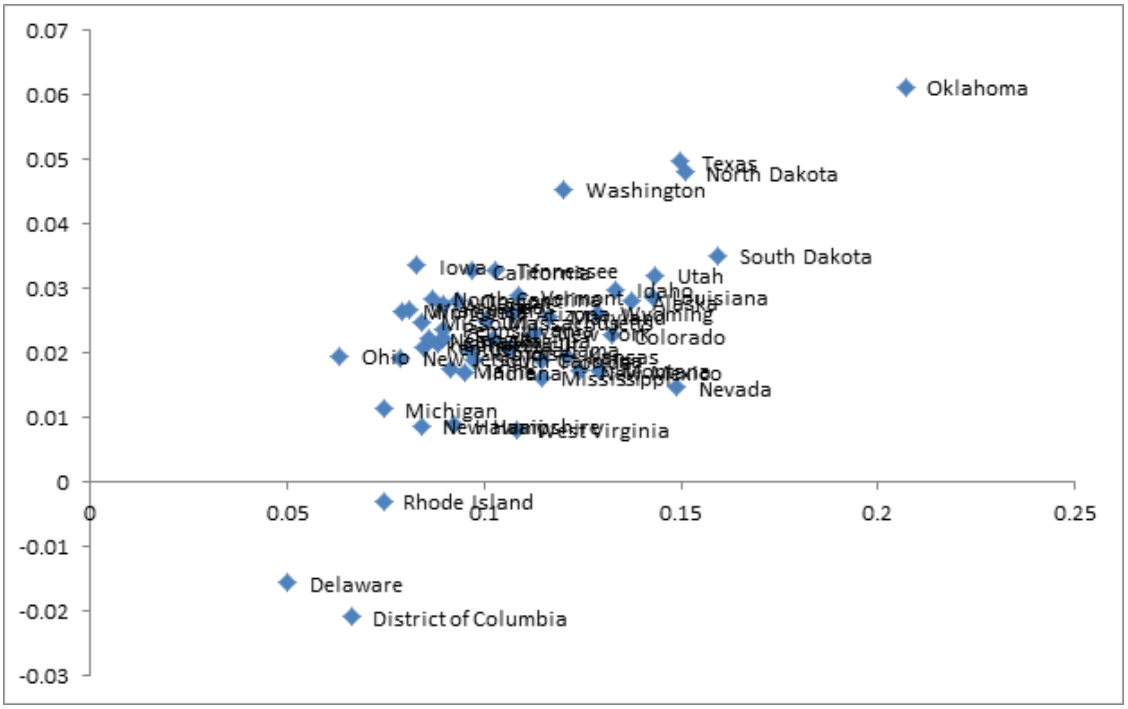
Source: Statistics computed from the Revenue enhanced LBD subset 1996-2000, 2003-2013. Reported are estimated effects of linear probability models on state effects.

Figure B10 State Rankings in Revenue HGF, Change from 96-06 to 07-13



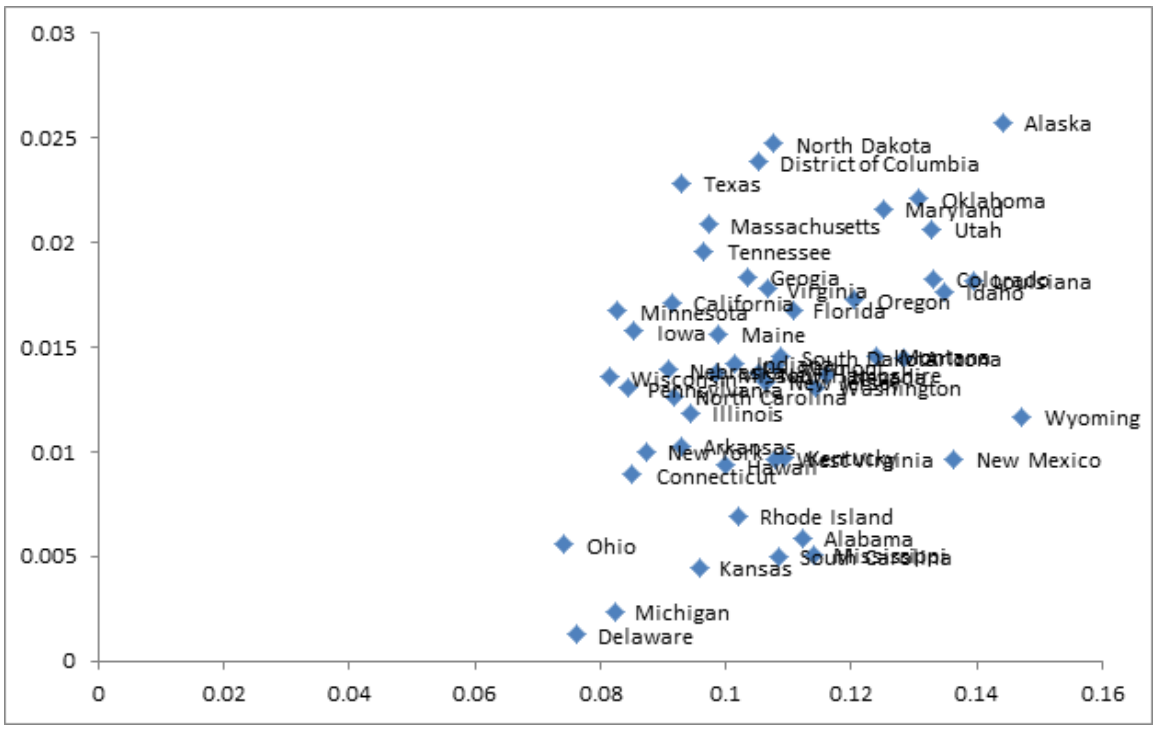
Source: Statistics computed from the Revenue enhanced LBD subset 1996-2000 and 2003-2013. The rankings for 1996-06 use the estimates from the 1996-06 period (except for 2001 and 2002) and the rankings of 2007-13 use the estimates from the 2007-13 period. Reported are estimated effects of linear probability models on state effects.

Figure B11 Revenue HGF versus Revenue Growth by State



Source: Statistics computed from the Revenue enhanced LBD subset 1996-2000, 2003-2013. Reported Revenue HGF are estimated effects of linear probability models on state effects. Mean Revenue Growth is revenue weighted mean revenue growth for firms.

Figure B12 Employment HGF versus Employment Growth by State



Source: Statistics computed from the Revenue enhanced LBD subset 1996-2000, 2003-2013. Reported Employment HGF are estimated effects of linear probability models on state effects. Mean Employment growth is employment weighted mean employment growth for firms.

Table B1: Correlations of High Growth Industry Effects with Summary Measures of First, Second and Third Moments of Industry Distributions

	Rev (GR)	Rev (HG)	Emp (GR)	Emp (HG)	Rev (HD)	Emp (HD)	Rev (90-10)	Emp (90-10)	Rev (Skew)	Emp (Skew)
Rev (GR)	1.00	0.52	0.39	0.06	-0.40	-0.16	-0.02	-0.07	0.40	0.16
Rev (HG)		1.00	0.28	0.47	0.43	0.36	0.76	0.44	0.15	0.15
Emp (GR)			1.00	0.52	-0.05	0.03	0.08	0.25	0.24	0.54
Emp (HG)				1.00	0.44	0.81	0.54	0.92	-0.02	0.35
Rev (HD)					1.00	0.53	0.81	0.52	-0.44	-0.02
Emp (HD)						1.00	0.57	0.94	-0.15	-0.07
Rev (90-10)							1.00	0.61	-0.24	0.05
Emp (90-10)								1.00	-0.11	0.13
Rev (Skew)									1.00	0.16
Emp (Skew)										1.00

Note: Rev=Revenue, Emp=Employment, GR= net growth, HG=high growth industry effect, HD=high decline industry effect, 90-10=activity weighted 90-10 differential (employment weights for Emp and revenue weights for Rev). Skew=(90-50)-(50-10) (activity weighted).

Table B2: Correlations of High Growth State Effects with Summary Measures of First, Second and Third Moments of State Distributions

	Rev (GR)	Rev (HG)	Emp (GR)	Emp (HG)	Rev (HD)	Emp (HD)	Rev (90-10)	Emp (90-10)	Rev (Skew)	Emp (Skew)
Rev (GR)	1.00	0.64	0.46	0.21	-0.36	0.06	0.20	0.17	0.61	0.15
Rev (HG)		1.00	0.45	0.66	0.37	0.54	0.82	0.63	0.43	0.31
Emp (GR)			1.00	0.33	-0.05	0.02	0.20	0.22	0.28	0.61
Emp (HG)				1.00	0.58	0.91	0.75	0.98	0.07	0.35
Rev (HD)					1.00	0.63	0.79	0.60	-0.46	0.22
Emp (HD)						1.00	0.73	0.97	-0.01	0.05
Rev (90-10)							1.00	0.77	0.01	0.25
Emp (90-10)								1.00	0.05	0.19
Rev (Skew)									1.00	0.01
Emp (Skew)										1.00

Note: Rev=Revenue, Emp=Employment, GR= net growth, HG=high growth industry effect, HD=high decline industry effect, 90-10=activity weighted 90-10 differential (employment weights for Emp and revenue weights for Rev). Skew=(90-50)-(50-10) (activity weighted).