EPOP
Understanding Entrepreneurship in the US

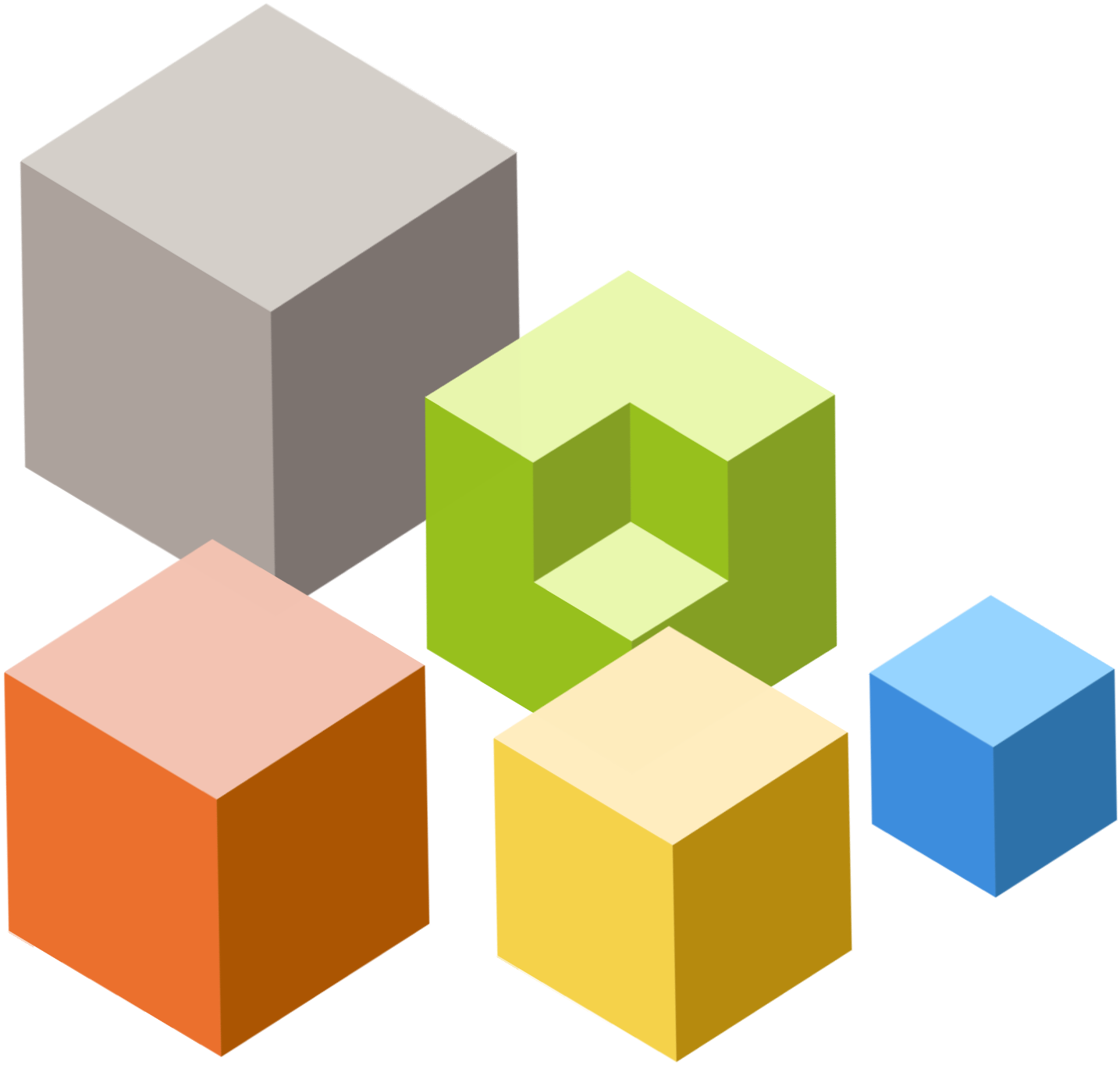
Entrepreneurship in the Population Survey

EPOP: 2023 Small Area Estimation Methodology Report

ISSUED DATE:
February 14, 2024

Created by
NORC at the University of Chicago
55 East Monroe Street, 30th Floor
Chicago, IL 60603
(312) 759-4000 Main
(312) 759-4004 Fax

Point of Contact
The NORC EPOP Research Team
EPOPresearch@norc.org



The Entrepreneurship in the Population Survey Project is being conducted by researchers at NORC at the University of Chicago with funding from a grant from the Ewing Marion Kauffman Foundation. Questions about this research project should be directed to EPOPresearch@norc.org.

The full title of the survey is "The Entrepreneurship in the Population Survey" and the abbreviation is EPOP Survey. In referencing the project or document, follow these standards:

Full Project Title:	The Entrepreneurship in the Population Survey Project: 2023
Project Abbreviation:	EPOP:2023
Full Report Title:	Entrepreneurship in the Population Survey Small Area Estimation Methodology Report: 2023
Report Abbreviation:	EPOP Small Area Estimation Methodology:2023
Citation:	"Entrepreneurship in the Population Survey Small Area Estimation Methodology: 2023." NORC at the University of Chicago. February 14, 2024. https://EPOP.norc.org."

TABLE OF CONTENTS

1. Introduction

This document provides a description of the small area estimation (SAE) methodology used to construct key entrepreneurship indicators using data from the second year of the Entrepreneurship in the Population (EPOP:2023) Survey, in conjunction with publicly available data sources. The use of SAE provides more precise estimates for rarer populations than what could be achieved using survey data alone. For reviews of SAE, see Rao and Molina (2015), Erciulescu et al. (2021), or Pfeffermann (2013).

All SAE estimates are available on the SAE page on the EPOP project website at: https://epop.norc.org/us/en/epop/about-the-study/small-area-estimates.html.

SAE estimates have also been incorporated into EPOP data dashboards: https://epop.norc.org/us/en/epop/researchers/interactive-data.html.

2. Overview of Small Area Estimation in EPOP Year 2

The SAE program based on EPOP:2023 data focused on **Entrepreneurial Activity Models.** SAE models were aimed at estimating the prevalence of individuals participating in an entrepreneurial activity, either overall or among people of a particular race or gender. For instance, an estimate might answer the question: what is the proportion of nascent business owners among Hispanics in Illinois?

SAE estimates were constructed for the following entrepreneurial activities:

1. Current business ownership
2. Current freelancing
3. Nascent entrepreneurship
4. Former business ownership
5. Former freelancing
6. Withdrawn entrepreneurship
7. Non-entrepreneurship (has never considered starting a business)
8. Gig work

Note that these entrepreneurial activities are not mutually exclusive, and any given individual can participate in more than one activity. For a more complete description of the definitions of these, refer to the EPOP:2023 Methodology Report.

Prevalence estimates were produced for all 50 states, the District of Columbia, and the top 50 most populated Metropolitan Statistical Areas (MSAs) in the US. Estimates were also produced for these geographies for individual race/ethnic groups (Hispanic, Non-Hispanic Black, Non-Hispanic White, and all other), and by gender (male/female).

3. ESTIMATION METHODOLOGY

EPOP:2023 small area estimates used the Fay-Herriot model (FH, Fay and Herriot, 1979), which is also used for the official estimation of proportions of children in poverty at the state and county level by the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) Program, among many other applications (Bell et al. 2016).

The FH model can be expressed as:

$$\widehat{Y}_i = \theta_i + e_i$$

$$\theta_i = x_i'\beta + u_i.$$

Above, $\widehat{Y}_i$ is the direct survey estimator of the quantity of interest for domain i , where there are $i = 1, .., m$ domains of interest, usually referred to as small areas (even though some of them are potentially large). The random variable e_i is the sampling error for domain i , x_i' is the vector of explanatory variables, and u_i is the area random effect independent of e_i.

The first level of the model describes the uncertainty due to sampling. The variance of e_i is the direct estimator's sampling variance, usually assumed to be known for identifiability. In practice, this variance needs to be estimated from the microdata, and sometimes, the direct estimators of sampling variances are smoothed. Here, smoothing was only employed for areas with problematic direct sampling variance estimates as will be discussed in more detail subsequently. The second level of the FH model, often called the linking model, explains the relationship between the underlying population quantity of interest and the covariates used to describe it. The area random effect is often called the model error and attempts to capture what cannot be explained by the covariates.

In the setting of estimating the prevalence of an entrepreneurial activity, $\widehat{Y}_i$, is the direct survey-weighted estimator of this quantity at the level of aggregation of interest. The subscript i then indexes the 50 states and DC, the 50 largest MSAs, or the-cross classification of these geographic areas with race/ethnicity or gender. The vector of covariates, x_i' , are drawn from various public sources described in Section 4.

The FH model, like other similar area-level models, yields model predictions that are very similar to the corresponding direct estimators for domains with large sample sizes. Hence, the covariates from auxiliary data play a more prominent role in areas with small sample sizes but do not substantially change the estimates for domains with very large sample sizes.

Models were fit using the ***emdi*** package in R (Kreutzmann et al. 2019). This package provides different options for estimating the model parameters, and Maximum Likelihood (ML) was used. The *fh* function provides Empirical Best Linear Unbiased Predictors (EBLUPs) and corresponding Mean Squared Error (MSE) estimates, and the *step* function can perform stepwise variable selection. The package uses the Prasad-Rao (1990) approximation to the MSE, which is second order unbiased. For more information about EBLUP estimation and the Fay-Herriot model, see Rao and Molina (2015).

OTHER TECHNICAL DETAILS

Estimates of Zero

In some unusual cases, the direct estimates of a proportion for a given type of entrepreneur and state or MSA were zero or one and were accompanied by direct sampling variance estimates of zero. While this issue is rare (occurring in only 93 of 5,656 direct estimates for the small areas of interest), these variance estimates are unrealistic and underestimate the true sampling variances.

These zero-sampling variance estimates were replaced with more conservative and realistic estimates using smoothing. More specifically, denote the design effect for any given area as $deff\_i$. This was estimated following the guidance of You and Hidiroglou (2023):

$$\widehat{deff}_i = \frac{\widehat{V}_i}{\hat{p}_i(1-\hat{p}_i)/n_i + \widehat{V}_i/n_i} \times \frac{n_i + 1}{n_i}$$

where $\widehat{V}_i$ is the direct estimate of the sampling variance, n_i the sample size, and $\hat{p}_i$ is the survey weighted estimator for the proportion of interest for the area. The sampling variances for domains with direct estimates of zero or one were then estimated by:

$$\widehat{V_i^{DEFF}} = \overline{deff} \times \frac{\bar{p}(1-\bar{p})}{n_i} \times \left(1 + \frac{1-\overline{deff}}{n_i}\right)^{-1}$$

where $\overline{deff}$ is the average estimated design effect across states or MSAs dropping observations of zero, and $\bar{p}$ is the average estimated proportion (where these averages were taken separately by demographic group when applicable). This formula was also recommended in You and Hidiroglou (2023) for use in SAE modeling, with empirical and simulation evidence of its good performance when used for the FH model. The corresponding entrepreneurship prevalence direct

estimates themselves were not changed and left at zero for fitting the models, but these new variance estimates represent more conservative and realistic values for the true sampling variance.

Additional Variance Smoothing

In some cases, the direct estimates of the sampling variance, though not zero, were extremely small, resulting in implausibly large effective sample sizes. This phenomenon is not unique to EPOP and was also observed in Rein et al. (2024) in American Community Survey (ACS) data. In these cases, the variance was again smoothed by computing and using an average estimated design effect. This smoothing was applied whenever the effective sample size for a given domain was estimated to be more than twice the actual sample size. These cases represented an additional 101 instances where smoothing occurred, beyond the 93 mentioned in the previous section.

Internal Consistency

Note that because state, state & gender, and state & race models were fitted separately, these models were not internally consistent in the sense that when multiplied by appropriate population totals, the state & gender and state & race totals for any given entrepreneurship type will not add up to the corresponding state totals. An analogous statement holds for the MSA model estimates. While fitting individual models at a lower level could resolve the internal consistency issue, the direct estimates at this more granular level of aggregation would have been based on very small sample sizes and therefore been more unstable.

4. Covariate Selection

Covariate Sources

The covariates used in the modeling were obtained from a variety of sources as documented in Table 1.

Table 1: Covariate Sources

Data Source	Link	Notes
American Business Survey (ABS)	https://www.census.gov/programs-surveys/abs.html	Data was available by both geographical level (state or MSA) by gender and geographical level by race/ethnicity.

American Community Survey (ACS)	https://www.census.gov/programs-surveys/acs	Data was available by both geographical level (state or MSA) by gender and geographical level by race/ethnicity.
Business Dynamic Statistics (BDS)	https://www.census.gov/programs-surveys/bds.html	Data were available at either of the geographic levels of interest (state or MSA).
Nonemployer Statistics (NES)	https://www.census.gov/programs-surveys/nonemployer-statistics.html	Data was available by both geographical level (state or MSA) by gender and geographical level by race/ethnicity.
Quarterly Workforce Indicators (QWI)	https://www.census.gov/data/developers/data-sets/qwi.html	Data was available by both geographical level (state or MSA) by gender and geographical level by race/ethnicity.
Kauffman Indicators of Entrepreneurship (KIE)	https://indicators.kauffman.org	Data was only used at the state level due to missingness at lower levels of aggregations and to reduce measurement error in the covariates error.
Internal Revenue Service (IRS)	https://www.irs.gov/statistics/soi-tax-stats-data-by-geographic-area	Tax summaries. Data were available at either of the geographic levels of interest (state or MSA).

The most recent available data for each data source was used, while taking the extent of missing cells into consideration. Data sources with substantial missingness were excluded. ACS 5-year estimates were used because of the lower variance relative to ACS 1-year estimates, which reduces potential problems of measurement error in the covariates (Bell et al. 2019).

The vintages used for the other datasets are as follows: ABS (2020), BDS (2021), QWI (2020-21), NES (2019), KIE (2021), IRS (2020). Missing cells were imputed by replacing them with the corresponding average across MSAs or states for a given demographic group.

A full listing of all potential covariates and those that were selected is included in Appendix A.

Covariate Selection Methodology

A stepwise selection algorithm was employed to identify the most promising covariates, using the R StepReg package (Li et al., 2022). This procedure assumes a simple linear regression model holds and is a good way to pre-screen covariates, but it is known that the models selected for such simplified models may not yield optimal results for the FH model (Lahiri and Suntornchost, 2014). Therefore, after pre-screening using ***StepReg,*** stepwise selection under the full FH model using the ***emdi*** package and used forward selection under the BIC criteria. All models included an intercept. In addition, for all models that were specific to race/ethnicity or gender groups, fixed effects were included for race or gender categories to better capture differences among the groups and to protect against potential racial or gender biases in the

covariates. Using the ***emdi*** package, the fixed effects were forced into the model, so that the stepwise regression found the best additional predictors given the fixed effects are included.

5. Data Suppression

Following the modeling exercise, some model estimates were suppressed due to their high uncertainty. This process mirrored the standard used by the U.S. Census Bureau's American Community Survey (ACS) to suppress estimates with coefficients of variation exceeding 0.61. Larger CVs imply 90% confidence intervals constructed from the published estimates and associated mean squared errors (MSEs) would fall outside the interval (0,1). This rule led to suppression of only 6 estimates for EPOP:2023.

6. Estimation Results

The use of SAE modeling resulted in large decreases in uncertainty measures. In most cases, the MSEs of the FH model were smaller than the estimated sampling variances of the direct estimators. In cases where this was not true, the differences were very small. Table 2 below shows the median percentage decrease of the model MSEs relative to the (possibly smoothed) sampling variance estimates. The median decreases for a given variable and level of stratification ranged from 30% to 83%, which show the SAE efforts yielded great benefits.

Table 2: Median Percentage Decrease in MSEs (Entrepreneurial Activity Models)

Entrepreneurial Activity Group	MSA	State	MSA & Gender	State & Gender	MSA & Race	State & Race
Current Entrepreneur	63%	64%	78%	77%	80%	75%
Current Freelancer	60%	70%	58%	65%	68%	80%
Former Entrepreneur	79%	69%	72%	76%	70%	71%
Former Freelancer	72%	74%	73%	65%	58%	75%
General Population	83%	65%	62%	67%	73%	56%
Gig Work	30%	43%	47%	53%	58%	58%
Nascent	56%	44%	73%	47%	57%	64%
Withdrawn	71%	73%	56%	55%	67%	68%

Note: Cells show the median percentage decrease of the MSEs of the model predictors relative to the (possibly smoothed) sampling variance estimate of the direct estimator of the proportion of entrepreneurs for each category and level of aggregation.

REFERENCES

Bell, W. R., H.-C. Chung, Datta, G. S., and Franco, C. (2019). Measurement Error in Small Area Estimation: Functional Versus Structural Versus Naive Models. *Survey Methodology*, 45, 61–80. https://www150.statcan.gc.ca/n1/pub/12-001-x/2019001/article/00005-eng.htm

Bell, W. R., Basel, W. W. and Maples, J. J. (2016). An Overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates Program. *In Analysis of Poverty Data by Small Area Estimation* (ed. M. Pratesi), Ch. 19. Chichester: Wiley.

Erciulescu, A., Franco, C., and Lahiri, P. (2021). Use of Administrative Records in Small Area Estimation. *Administrative Records for Survey Methodology*. New York: Wiley.

Fay, R.E., and Herriot, R.A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, Vol. 74 (366), pp. 269-277.

Kreutzmann A, Pannier S, Rojas-Perilla N, Schmid T, Templ M, Tzavidis N (2019). "The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators." Journal of Statistical Software, 91(7), 1–33. doi:10.18637/jss.v091.i07.

Lahiri, P., and Suntornchost, J. (2015). Variable Selection for Linear Mixed Models with Applications in Small Area Estimation. *Sankhya B*, 77, 312-320. https://doi.org/10.1007/s13571-015-0096-0.

Li, J., La, X, Cheng, K., and Liu, W. (2022). StepReg: Stepwise Regression Analysis. Version 1.4.3. https://cran.r-project.org/web//packages/StepReg/StepReg.pdf

Pfeffermann, D., (2013). New Important Developments in Small Area Estimation. *Statistical Scienc*e, Vol. 28, No. 1, 40−68.

Prasad, N. N., & Rao, J. N. (1990). The estimation of the mean squared error of small-area estimators. Journal of the American statistical association, 85(409), 163-171.

R Core Team (2021). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rein, D. B., Franco, C., Reed, N. S., Herring-Nathan, E. R., Lamuda, P. A., Alfaro Hudak, K. M., Hu, W., Hartzman, A. J., Hite, K. R., Wittenborn, J. S. (2024). The Prevalence of Bilateral Hearing Loss in the United States in 2019: A Small Area Estimation Approach to Obtain National, State, and County Level Estimates by Demographic Subgroup. To appear in The Lancet Regional Health - Americas.

Rao, J. N., and Molina, I. (2015). Small area estimation. New York, NY: John Wiley.

Stan Development Team (2018). RStan: The R Interface to Stan. R Package Version 2.17.3. http://mc-stan.org/

Stan Development Team (2022). Stan Modeling Language Users Guide and Reference Manual. Version 2.31.0. http://mc-stan.org/

You, Y., & Hidiroglou, M. (2023). Application of Sampling Variance Smoothing Methods for Small Area Proportion Estimation. Journal of Official Statistics, 39(4), 571-590.

APPENDIX A: COVARIATES USED IN ESTIMATION

APPENDIX A – COVARIATES SELECTED BY MODEL

Table 1: Covariates Selected by Model

Geographic Level	Entrepreneurial Activity	Covariates
MSA	Current Business Ownership	Employee establishment age (left censored) Federal government workers Employee firm age (2020 Quarter 4, age = 2 to 3 years) Employee firm size (2020 Quarter 4, size = 250-499 employees) Employee firm size (2021 Quarter 3, size = 20-49 employees) Employee firm age (2021 Quarter 1, age = 2 to 3 years)
MSA	Former Freelancing	Employee establishment age (6 to 10 years) Employed residents per working age population Employee firm size (2020 Quarter 4, size = 20-49 employees) Employee firm size (2021 Quarter 3, size = 250-499 employees)
MSA	Non-entrepreneurship	Employee establishment age (left censored) Federal government workers
MSA	Current Freelancing	Employee establishment age (left censored) Employed residents per working age population
MSA	Gig Work	Employee establishment age (0 to 1 year) Employed residents per working age population
MSA	Nascent Entrepreneurship	Employee establishment age (left censored) Employed residents per working age population Local government workers Business or professional income tax returns Employee firm age (2020 Quarter 4, age = 2 to 3 years)
MSA	Withdrawn Entrepreneurship	Employed residents per working age population State government workers
MSA	Former Business Ownership	Employee establishment age (left censored)
MSA x Gender	Current Business Ownership	Employee establishment age (left censored)
MSA x Gender	Former Freelancing	Employee establishment age (6 to 10 years) Employee firm size (2020 Quarter 4, size = 20-49 employees) Employee firm size (2021 Quarter 3, size = 250-499 employees)
MSA x Gender	Non-entrepreneurship	Employee establishment age (left censored)

Geographic Level	Entrepreneurial Activity	Covariates
MSA x Gender	Current Freelancing	Employee establishment age (left censored) Employed residents per working age population
MSA x Gender	Gig Work	Employee establishment age (0 to 1 year) Employed residents per working age population
MSA x Gender	Nascent Entrepreneurship	Employee establishment age (left censored) Employed residents per working age population Local government workers Non-employers per working age population Business or professional income tax returns Employee firm age (2020 Quarter 4, age = 0 to 1 year) Employee firm age (2020 Quarter 4, age = 2 to 3 years) Self-employed in own incorporated business workers Self-employed in own not incorporated business workers
MSA x Gender	Withdrawn Entrepreneurship	Employed residents per working age population
MSA x Gender	Former Business Ownership	Employee establishment age (left censored)
MSA x Race	Current Business Ownership	Employee firm age (2020 Quarter 4, age = 2 to 3 years) Employee firm size (2020 Quarter 4, size = 20-49 employees)
MSA x Race	Former Freelancing	Self-employed in own incorporated business workers
MSA x Race	Non-entrepreneurship	Employee firm age (2020 Quarter 4, age = 0 to 1 year) Employee firm size (2020 Quarter 4, size = 500+ employees) Employee firm age (2021 Quarter 3, age = 2 to 3 years)
MSA x Race	Current Freelancing	Self-employed in own incorporated business workers
MSA x Race	Gig Work	Employee firm size (2021 Quarter 2, size = 20-49 employees)
MSA x Race	Nascent Entrepreneurship	No additional covariates
MSA x Race	Withdrawn Entrepreneurship	Private not for profit wage and salary workers Self-employed in own incorporated business workers
MSA x Race	Former Business Ownership	Employee firm age (2021 Quarter 3, age = 2 to 3 years)
State	Current Business Ownership	Employee establishment age (11 to 15 years) Non-employers per working age population Employee firm size (2021 Quarter 2, size = 20-49 employees) Rate of new entrepreneurs State government workers
State	Former Freelancing	Employee establishment age (left censored) Employee establishment age (0 years) Federal government workers

Geographic Level	Entrepreneurial Activity	Covariates
State	Non-entrepreneurship	Employee establishment age (left censored) Federal government workers Self-employed in own not incorporated business workers
State	Current Freelancing	Self-employed in own not incorporated business workers Employee establishment age (4 years) Non-employers per working age population Business or professional income tax returns Unpaid family workers
State	Gig Work	Employed residents per working age population Business or professional income tax returns Employee firm size (2021 Quarter 2, size = 20-49 employees) Private not for profit wage and salary workers State government workers
State	Nascent Entrepreneurship	Employee establishment age (11 to 15 years) Business or professional income tax returns Employee firm size (2021 Quarter 2, size = 20-49 employees)
State	Withdrawn Entrepreneurship	Employee firm size (2021 Quarter 1, size = 20-49 employees)
State	Former Business Ownership	Employee establishment age (left censored) Percentage of tax returns with self-employment tax Self-employed in own not incorporated business workers
State x Gender	Current Business Ownership	Employee establishment age (11 to 15 years) Non-employers per working age population State government workers Kaufman summary index
State x Gender	Former Freelancing	Employee establishment age (16 to 20 years) Federal government workers Rate of new entrepreneurs
State x Gender	Non-entrepreneurship	Employee establishment age (left censored) Federal government workers Self-employed in own not incorporated business workers
State x Gender	Current Freelancing	Rate of new employer business actualization Employee establishment age (4 years) Employed residents per working age population
State x Gender	Gig Work	Employed residents per working age population Business or professional income tax returns Employee firm size (2021 Quarter 2, size = 20-49 employees)
State x Gender	Nascent Entrepreneurship	Employee establishment age (11 to 15 years) Business or professional income tax returns

Geographic Level	Entrepreneurial Activity	Covariates
		Employee firm size (2021 Quarter 2, size = 20-49 employees)
State x Gender	Withdrawn Entrepreneurship	No additional covariates
State x Gender	Former Business Ownership	Employee firm age (2021 Quarter 3, age = 0 to 1 year) Self-employed in own not incorporated business workers
State x Race	Current Business Ownership	Employee firm age (2021 Quarter 3, age = 0 to 1 year) Unpaid family workers
State x Race	Former Freelancing	Federal government workers Self-employed in own not incorporated business workers
State x Race	Non-entrepreneurship	Employee establishment age (11 to 15 years) Unpaid family workers
State x Race	Current Freelancing	Rate of new employer business actualization Employee firm age (2021 Quarter 3, age = 0 to 1 year)
State x Race	Gig Work	Employee firm age (2021 Quarter 1, age = 0 to 1 year) Employee firm size (2021 Quarter 1, size = 20-49 employees) Employee firm age (2021 Quarter 3, age = 0 to 1 year)
State x Race	Nascent Entrepreneurship	No additional covariates
State x Race	Withdrawn Entrepreneurship	No additional covariates
State x Race	Former Business Ownership	Employee establishment age (11 to 15 years) Employee firm age (2021 Quarter 3, age = 0 to 1 year) Self-employed in own incorporated business workers Startup early survival rate