Entrepreneurship in the Population Survey

EPOP: 2022 Small Area Estimation Methodology Report

ISSUED DATE:
September 1, 2023

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: 2022
Project Abbreviation:	EPOP:2022
Full Report Title:	Entrepreneurship in the Population Survey Small Area Estimation Methodology Report: 2022
Report Abbreviation:	EPOP Small Area Estimation Methodology:2022
Citation:	"Entrepreneurship in the Population Survey Small Area Estimation Methodology: 2022." NORC at the University of Chicago. September 1, 2023. https://EPOP.norc.org."

Table of Contents

1. Introduction

This document provides a description of the methodology underpinning estimates using small area estimation (SAE) that were constructed for key entrepreneurship indicators using data from Year 1 of the Entrepreneurship in the Population Survey (EPOP) in conjunction with publicly available data sources. The use of SAE provides more precise estimates for rarer populations that would be achieved using survey data alone.

All SAE estimates are available on the EPOP project website at EPOP.norc.org. SAE estimates are 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 1

There were two general groups of models conducted for SAE using EPOP Year 1 data, where the models differed based on the estimand of interest:

1. **Entrepreneurial Activity Models.** The first group of models estimated the prevalence of individuals participating in an entrepreneurial activity, either overall or among people of a particular race or gender. This could also be thought of as the conditional probability of being in a particular type of entrepreneurship category given race or gender. For instance, this quantity might answer the question: what is the proportion of nascent business owners among Hispanics in Illinois?

2. **Demographic Composition Models.** The second group of models focused on estimating the demographic composition of individuals participating in a given entrepreneurial activity. This can also be thought of as the probability of being of a particular race or gender conditional on being a particular tyle of entrepreneur. For instance, this estimate my answer the question: what is the proportion of New York current business owners that are female?

These 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 1 entrepreneurial activity. For a more complete description of the definitions of these entrepreneurial activities, refer the EPOP Year 1 Methodology Report.

Estimates of the prevalence of entrepreneurial activities were created for all 50 states, DC, and the top 50 most populated Metropolitan Statistical Areas (MSAs) in the US. Both the prevalence estimates and the estimates of the demographic composition of types of entrepreneurship were created for individual race/ethnic groups (Hispanic, Non-Hispanic Black, and all other), and by gender.

3. Estimation Methodology

The methodology employed for SAE differed according to the type of estimand being considered. Below, we describe separately the methodology used for estimating the prevalence of entrepreneurial activities as well as for estimating the demographic composition of these entrepreneurial groups.

Entrepreneurial Activity Models

When estimating the prevalence of given entrepreneurial activities, we used the most-established and widely used model in small area estimation, the Fay-Herriot model (FH, Fay and Herriot, 1979). The FH model is used in the important application of 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 2015).

Generally, 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, since we do not observe the domain's quantity of interest but use a noisy survey estimate of it in our models. The variance of e_i is the direct estimator's sampling variance, usually assumed known for identifiability. In practice, this variance needs to be estimated from the microdata, and sometimes, the direct estimators of sampling variances are smoothed, though we didn't do this here. 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 below in Section 4.

The FH model, and other similar area-level models, yield 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 large sample sizes.

Models were initially fit using the sae package in R (Molina and Marhuenda, 2020). This package provides different options for estimating the parameters, and we used the default which is Restricted Maximum Likelihood (REML). We used this frequentist implementation for a speedy way to do stepwise variable selection using the Fay-Herriot model by using the results of the model fitting to program the stepwise algorithm. We used Bayesian techniques to fit the final models. The sae package uses the Prasad-Rao (1990) approximation to the mean squared error (MSE), which is second order unbiased (Rao and Molina, 2015).

A Bayesian implementation of the FH model can be fit via software like Stan (Stan Development Team, 2022), and packages like rstan (R Development Team, 2018). For our analysis, we used rstan as a convenient interface. The Stan software enables the user to compute approximations to the posterior distribution of the parameters of a given model, and to calculate estimates of the posterior mean and variance based on Markov Chain Monte Carlo (MCMC). For this project, we used diffuse priors, which attempt to assume little or no prior knowledge about the model parameters, other than obvious constraints such as having variances be positive. We used a diffuse uniform prior on the standard deviation of the random effects, except when it resulted in lack of convergence. In such cases, we used a diffuse gamma prior on the precision parameter. For the regression coefficients, we used a diffuse normal prior.

The use of a Bayesian implementation for fitting the final models provides two benefits for the current setting. First, in some cases the frequentist implementation gave a model variance estimate of zero (when the true parameter is near zero). A model variance estimate of zero is not a realistic value and implies the resulting model estimates are fully synthetic and would not necessarily be close to the direct estimators for large sample sizes. A Bayesian approach can remedy this and has been previously used to address this issue for SAIPE estimates of school-aged children in poverty (Bell et al., 2015).

In addition, the Bayesian approach also can produce demographic composition estimates with a relatively minor modification to the Stan modeling code. This obviates the need to perform a new search of covariates and to fit entirely new models for the demographic composition estimates. Furthermore, deriving the demographic composition estimates from the models for entrepreneurship activities provides an approximate internal consistency among the two types of estimates. It should be noted that both the estimates and measures of uncertainty were very similar, except in the cases where the frequentist implementation resulted in a model variance of zero.

Demographic Composition Models

When estimating the proportion of people identifying as a given race/ethnicity or gender among those engaged in an entrepreneurial activity, we compute estimates from the models discussed previously. Using Bayes' formula, the estimand of interest can be expressed as functions of the probabilities estimated by the corresponding Fay Herriot models above, which modeled the probability of performing a given entrepreneurial activity conditional on either race/ethnicity or gender. For instance, suppose that for a given domain we are interested in the probability being from a particular demographic group given we are a particular type of entrepreneur. Denote this probability as $P(D_1|E)$. Then according to Bayes' formula, we can express this probability as follows:

$$P(D_1|E) = P(E|D_1)P(D_1)/P(E)$$

Where $P(D_1)$ is the probability of belonging to the demographic group within the domain, $P(E|D_1)$ is the probability of being an entrepreneur given one belongs to the demographic group in question, and $P(E)$ is the probability of being a particular type of entrepreneur. Note that $P(E|D_1)$ was already obtained in the models described in the previous section. $P(D_1)$ can be obtained from ACS population data. For $P(E)$, in order to obtain them from the same model as $P(E|D_1)$, hence ensuring sensible estimates that add up to one, we used the total probability formula:

$$P(E) = \sum P(E|D_i)\, P(D_i)$$

With these expressions, the estimates of the mutually exclusive and exhaustive conditional probabilities of belonging to a race automatically add up to one. Hence, the models we described in the previous section for MSA and gender, MSA and race, state and race, and state and gender automatically could produce estimands for the demographic composition models with minor modifications which instructed Stan to also compute the functions described above with the output.

Once we have a function that expresses the new probabilities (probability of belonging to a demographic group conditional on entrepreneurial activity) in terms of the already modeled probabilities (probability of and entrepreneurial activity conditional on belonging to a demographic group), this function can be applied in each iteration of the MCMC replications for the models for entrepreneur type described in the previous section. We can then approximate the posterior distribution of the new proportions of interest via standard MCMC methods and software.

One assumption made in this approach is that the proportions of the population belonging to a particular race or gender, within a given state or MSA, are known and not subject to sampling variability. This assumption should be approximately realistic as these population estimates were obtained from the ACS 5-year estimates for 2019 which have a very large sample sizes, and hence their sampling variability should be negligible compared to those of the direct estimates used in the modeling.

In a small number of cases (a total of 10 estimates coming from 4 distinct models), the approach described in the preceding paragraph resulted in extremely high variances, likely due to numerical issues. For these cases, we fit new FH models on the direct estimates of the probabilities of being of a given race given one belongs to an entrepreneurship class directly, using new covariates that were more likely to be predictive for these probabilities. The models that used this different methodology included withdrawn entrepreneurship for state by gender, for state by race, and for MSA by race, and current business ownership for MSA by race. Note these four new models were used to produce all of the corresponding estimates for the probability of belonging to a particular race or gender given one belongs to a particular entrepreneur class, not just to produce estimates for the cases that had shown extremely high variances. This was for consistency across areas and because the numerical instability could have potentially affected other estimates in a less obvious ways. However, the original models were retained for estimating the probability of being a specific type of entrepreneur given one belongs to a demographic group within a domain. The original models showed good behavior for these and there was no reason to change the methodology.

Other Technical Details

Direct Estimate of Zero

In some unusual cases, the direct estimate of a proportion for a given type of entrepreneur and state or MSA were zero and were accompanied by direct sampling variance estimates of zero.

This is of course not a realistic estimate, as these zero estimates are associated with tiny sample sizes. Hence, we sought a better estimate of the sampling variance for our model fitting.

The effect of underestimating the sampling variance in models such as the Fay-Herriot model is to place undue weight on these noisy estimates both in the parameter estimation and in the particular area's modeled estimates. In fact, Bell (2008) showed that under-estimating the sampling variance can yield more severe problems than over-estimating. Hence, we replaced the zero-sampling variance estimates with more conservative estimates. To compute these, we estimated the design effect $deff_i$ for these areas by taking an average across the states or MSAs for that demographic group. We then used this to estimate the effective sample size $neff_i = n_i/deff_i$.

Finally, we substituted the average proportion, $\bar{p}$, for the observed survey-weighted proportion for the variable in question within a given demographic group across states or MSAs solely for the purpose of variance estimation. We then calculated variance using the following formula:

$$Var(p_i) = [\overline{p_\iota}(1 - \overline{p_\iota})]/neff_i.$$

The entrepreneurship prevalence estimates themselves were not changed and left at zero, but these new variance estimates provide more conservative realistic estimates of variance.

Internal Consistency

Note that because state, state and gender, and state and race models were fitted separately, they are not internally consistent in the sense that when multiplied by appropriate population totals, the state and gender and state and 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 model 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

The number of possible covariates we gathered was relatively large, and only a subset of potential covariates was used in the estimation. Both types of estimates described above used the same covariate selection process.

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 (see Bell et al. 2019 for more information about measurement error in small area estimation).
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 missing cells into consideration. Data sources with extensive missingness were excluded. We used ACS 2020 5-year data for the covariates for the state and MSA models, and 2015 5-year data for the models by geography and race/ethnicity. The 5-year estimates were collected over a period of time. The primary advantage of using multi-year estimates is their lower variances compared to estimates for just one year, which reduces problems due to measurement error in the covariates.

The vintages used for the other datasets are as follows: ABS (2019), BDS (2019), QWI (2015), NES (2018), KIE (2021), IRS (2019). Missing cells were imputed to have a complete set of covariates, replacing the missing cell with the corresponding average across MSAs or states for a given geographic group.

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

Covariate Selection Methodology

To identify the most promising covariates, we first used stepwise regression fit using the R StepReg package (Li et al., 2022). This procedure assumes a simplified version of the model with no random effects. That is, this procedure fits models under a simple linear regression. This is a good way to pre-screen covariates, but it is known that the models selected in the absence of random effects may not be best in the presence of random effects (Lahiri and Suntornchost, 2014). However, using the full-Fay Herriot model with a large number of covariates typically leads to a lack of convergence. Hence, after pre-screening using StepReg we performed a backward selection using the full Fay-Herriot model where we removed insignificant covariates one by one in steps, provided that the BIC decreased. At each step, the covariate with the highest *p*-value was removed. Sometimes, after the initial StepReg stepwise regression, the Fay-Herriot model did not converge with the initial group of variables selected. In such cases, we dropped some additional covariates, those with the highest p-values, to be able to continue with the backwards selection under the Fay-Herriot model assumptions.

All models included an intercept. In addition, for all models that were specific to race/ethnicity or gender groups, we included fixed effects or intercepts for race or gender categories to better capture differences among the groups.

5. Data Suppression

Following the estimation, some estimates were suppressed on the basis of reliability. We followed the standard used by the US Census Bureau's American Community Survey, to suppress estimates with coefficients of variation surpass in excess of 0.60. We also suppressed a small number of estimates where the SAE estimate was negative.

Table 2 summarizes how many suppressions were made, where some rows show both the proportion of participating in each entrepreneurial activity for people of a given race (p1), and the proportion of people of a particular race among those who participate in a given entrepreneurial activity (p2).

Table 2: Prevalence of Suppression Across Estimates

Type of Estimate	CV > 0.61 (p1/p2)	Percentage of zero estimates	Number of negative estimates	Total estimates computed
State - Overall	0	0	0	408
MSA - Overall	0	0	0	400
State - Gender	<1% / 0	0	0	1,632
MSA - Gender	0 / <1%	0	0	1,600
State - Race	4% / 5%	0.08%	1/1	2,448
MSA - Race	8% / 12%	0 / 0.25%	0/3	2,400
TOTAL	4%	0.06%	5	8,888

6. ESTIMATION RESULTS

The use of SAE resulted in large decreases in uncertainty measures. In most cases, the mean squared error of the FH model estimates were smaller than the direct variances. In cases where this was not true, the differences were very small. Table 2 shows the median percentage decrease of the posterior variance of the model estimates relative to the variance of the direct estimator. The median decreases for a given variable and level of stratification ranged from 43% to 85% for the case of the probability of a given entrepreneurship type.

Table 2: Median Percentage Decrease in Variance of Estimates: Entrepreneurial Activity Models

Entrepreneurial Activity Group	MSA	State	MSA & Gender	State & Gender	MSA & Race	State & Race
Current Entrepreneur	72%	78%	75%	81%	60%	67%
Current Freelancer	76%	80%	82%	79%	57%	69%
Former Entrepreneur	80%	75%	79%	61%	70%	59%
Former Freelancer	76%	80%	76%	85%	58%	65%
General Population	76%	84%	75%	67%	43%	63%
Gig Work	78%	78%	70%	50%	55%	71%

Nascent	56%	64%	68%	80%	52%	66%
Withdrawn	82%	76%	62%	76%	74%	70%

Note: Cells show the median percentage decrease of the posterior variances of the model predictors relative to the direct variance estimate of the direct estimator of the proportion of entrepreneurs for each type of entrepreneur category and level of aggregation.

Table 3 shows the corresponding reduction in the posterior variance for the probability of being from a given race or gender among entrepreneurs ranged from 42-94%. This again suggests that the SAE modeling was successful in reducing measures of uncertainty and producing more stable estimates compared to the direct survey estimators.

Table 3: Median Percentage Decrease in Variance of Estimates: Entrepreneurial Activity Models

Entrepreneurial Activity Group	MSA	State	MSA & Gender	State & Gender	MSA & Race	State & Race
Current Entrepreneur	72%	78%	75%	81%	60%	67%
Current Freelancer	76%	80%	82%	79%	57%	69%
Former Entrepreneur	80%	75%	79%	61%	70%	59%
Former Freelancer	76%	80%	76%	85%	58%	65%
General Population	76%	84%	75%	67%	43%	63%
Gigwork	78%	78%	70%	50%	55%	71%
Nascent	56%	64%	68%	80%	52%	66%
Withdrawn	82%	76%	62%	76%	74%	70%

Note: Cells show the median percentage decrease of the posterior variances of the model predictors relative to the direct variance estimate of the direct estimator of the proportion of entrepreneurs for each type of entrepreneur category and level of aggregation.

Table 3 shows the corresponding reduction in the posterior variance for the probability of being from a given race or gender among entrepreneurs ranged from 42-94%. This again suggests that the SAE modeling was successful in reducing measures of uncertainty and producing more stable estimates compared to the direct survey estimators.

Reference

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.

Bell, W. R. (2008). Examining Sensitivity of Small Area Inferences to Uncertainty About Sampling Error Variances. *In Proceedings of the American Statistical Association*, Survey Research Methods Section, Volume 327, pp. 334.

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.

Franco, C., and Bell, W. R. (2022). Using American Community Survey Data to Improve Estimates from Smaller US Surveys through Bivariate Small Area Estimation Models. *Journal of Survey Statistics and Methodolog*y, 10, 1, 225-247

Franco, C., and Bell, W. R. (2015). Borrowing Information Over Time in Binomial/Logit Normal Models for Small Area Estimation. *Statistics in Transition (new series) and Survey Methodology, Joint Issue on Small Area Estimation*, 16, 563—584. http://stat.gov.pl/en/sit-en/issues-and-articles-sit/previous-issues/volume-16-number-4-december-2015/

Franco, C., and W. R. Bell (2013). Applying Bivariate Binomial/Logit Normal Models to Small Area Estimation. *Proceedings of the American Statistical Association*, pp. 690–702.

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

Molina, I. and Marhuena, Y. (2020). Sae: Small Area Estimation. R Package Version 1.3. https://cran.r-project.org/web/packages/sae/sae.pdf

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

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

Rao, J.N.K. (2014). Inferential issues in model-based small area estimation: some new developments. *Statistics in Transition New Series, 2015*, Vol. 16, No. 4, pp. 491–510.

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/

Yang, M., Breidt, J., Ganesh, N., Mulrow, E., Nathan, E. H., and Pineau, V. (2022). "Hybrid Methods for Combining Probability and Nonprobability Samples for Estimation," *Joint Statistical Meetings 2022 Proceedings (forthcoming).*

APPENDIX A: COVARIATES USED IN ESTIMATION

Table A – Covariates Selected by Model

Geographic Level	Entrepreneurial Activity	Covariates
MSA	Current Business Ownership	Employee establishment age (left censored) Employee firm age (Quarter 2, age = 11+ years) Employee firm size (Quarter 2, size = 500+ employees) Employee firm age (Quarter 1, age = 2 years) Employed resident rate Employee firm age (Quarter 1, age = 4 years)
MSA	Former Freelancing	Employee firm size (Quarter 3, size = 20-49 employees) Employee firm size (Quarter 4, size = 20-49 employees) Employee establishment age (6 to 10 years) Employee firm size (Quarter 2, size = 20-49 employees)
MSA	Non-entrepreneurship	Employee establishment age (left censored) Unpaid family workers employee firm size (Quarter 1, size = 20-49 employees) Employee firm size (Quarter 2, size = 250-499 employees) Employee firm size (Quarter 2, size = 500+ employees) Employee firm age (Quarter 2, age = 0 to 1 years) Employee establishment age (6 to 10 years) Employee firm size (Quarter 3, size = 250-499 employees)
MSA	Current Freelancing	Employee firm size (Quarter 3, size = 20-49 employees) Employee firm size (Quarter 4, size = 20-49 employees) Employee establishment age (6 to 10 years) Employee firm size (Quarter 2, size = 20-49 employees)
MSA	Gig Work	Employee firm size (Quarter 4, size = 20-49 employees)
MSA	Nascent Entrepreneurship	Business or professional income tax returns
MSA	Withdrawn Entrepreneurship	Local government workers employee firm age (Quarter 4, age = 6 to 10 years)
MSA	Former Business Ownership	Unpaid family workers

MSA x Gender	Current Business Ownership	Employee firm age (Quarter 1, age = 11+ years) Employee firm size (Quarter 1, size = 250-499 employees) Employee establishment age (0 years)
MSA x Gender	Former Freelancing	Self-employed: incorporated Employee firm size (Quarter 2, size = 500+ employees) employee firm size (Quarter 4, size = 0-19 employees) Employee firm size (Quarter 2, size = 250-499 employees)
MSA x Gender	Non-entrepreneurship	Private, not for profit, wage and salary workers Employee firm size (Quarter 3, size = 20-49 employees) Employee firm size (Quarter 3, size = 500+ employees) State government workers federal government workers Employee firm age (Quarter 2, age = 0 to 1 years) Privately held firm rate Employee establishment age (left censored) employee establishment age (6 to 10 years)
MSA x Gender	Current Freelancing	Employee firm age (Quarter 1, age = 0 to 1 years) Employee firm age (Quarter 1, age = 11+ years) Employee firm size (Quarter 3, size = 20-49 employees) Employee establishment age (left censored) Employee firm size (Quarter 2, size =500+ employees) Employee firm size (Quarter 4, size = 500+ employees) Employee firm size (Quarter 2, size = 20-49 employees)
MSA x Gender	Gig Work	Private, not for profit, wage and salary workers Employee firm size (Quarter 3, size = 20-49 employees)
MSA x Gender	Nascent Entrepreneurship	Self-employment tax returns Business or professional income tax returns Employee establishment age (left censored) Employee firm size (Quarter 1, size = 0-19 employees) Employee firm size (Quarter 2, size = 20-49 employees) Employee firm age (Quarter 1, age = 6 to 10 years) employee firm age (Quarter 1, age = 11+ years) Employee firm age (Quarter 2, age = 2 to 3 years) employee firm age (Quarter 4, age = 0 to 1 years)
MSA x Gender	Withdrawn Entrepreneurship	No additional covariates

MSA x Gender	Former Business Ownership	Employee firm size (Quarter 4, size = 20-49 employees) Employee firm size (Quarter 2, size = 250-499 employees) Local government workers
MSA x Race	Current Business Ownership	Self-employed: non-incorporated employee firm age (Quarter 2, age = 6 to 10 years) Employee firm size (Quarter 3, size = 500+ employees)
MSA x Race	Former Freelancing	Nonemployer rate employee firm size (Quarter 2, size = 500+ employees) Employee firm age (Quarter 2, age = 0 to 1 years) Employee firm age (Quarter 2, age = 11+ years) Employee firm age (Quarter 3, age = 11+ years)
MSA x Race	Non-entrepreneurship	Employee firm size (Quarter 1, size = 0-19 employees) Employee firm age (Quarter 4, age = 0 to 1 years)
MSA x Race	Current Freelancing	Employee firm age (Quarter 2, age = 11+ years) Employee establishment age (left censored)
MSA x Race	Gig Work	Employee firm size (Quarter 2, size = 500+ employees) Employee firm size (Quarter 3, size = 500+ employees) employee firm size (Quarter 2, size = 0-19 employees) Employee firm size (Quarter 3, size = 0-19 employees)
MSA x Race	Nascent Entrepreneurship	Employee firm size (Quarter 1, size = 20-49 employees)
MSA x Race	Withdrawn Entrepreneurship	Employee firm size (Quarter 3, size = 0-19 employees) Employee firm size (Quarter 4, size = 0-19 employees) Employee firm age (Quarter 3, age = 6 to 10 years) Employee firm age (Quarter 4, age = 6 to 10 years)
MSA x Race	Former Business Ownership	Employee firm size (Quarter 4, size = 250-499 employees)
State	Current Business Ownership	Employee establishment age (left censored) Unpaid family workers
State	Former Freelancing	Employee establishment age (21 to 25 years) employee establishment age (1 year)

State	Non-entrepreneurship	Employee establishment age (left censored) Employee establishment age (5 years) Unpaid family workers
State	Current Freelancing	Employee establishment age (2 years) Employed resident rate
State	Gig Work	Unpaid family workers employee establishment age (26+ years) Employee establishment age (2 years)
State	Nascent Entrepreneurship	Business or professional income tax returns employee establishment age (21 to 25 years) Unpaid family workers
State	Withdrawn Entrepreneurship	Federal government workers Self-employed: incorporated Employed resident rate Employee establishment age (11 to 15 years) Employee establishment age (left censored)
State	Former Business Ownership	Employee establishment age (21 to 25 years) Opportunity share of new New employer business actualization Employee establishment age (0 years) Employee firm size (Quarter 3, size = 250-499 employees) Employee firm size (Quarter 1, size = 250-499 employees) Employed resident rate
State x Gender	Current Business Ownership	Employee firm size (Quarter 1, size = 500+ employees) Employee firm size (Quarter 2, size = 500+ employees) Zindex Employed resident rate Effective sample size
State x Gender	Former Freelancing	Employee establishment age (1 year) Employee firm size (Quarter 3, size = 0-19 employees) Self-employed: incorporated Employee establishment age (21 to 25 years)

State x Gender	Non-entrepreneurship	Employee firm size (Quarter 1, size = 250-499 employees) Employee firm age (Quarter 3, age = 2 to 3 years) Self-employed: non-incorporated employee establishment age (11 to 15 years) New entrepreneur rate employee firm age (Quarter 2, age = 2 to 3 years) Employee firm size (Quarter 4, size = 250-499 employees) Employee firm size (Quarter 1, size = 500+ employees) Employee firm size (Quarter 1, size = 0-19 employees) Employee establishment age (1 year)
State x Gender	Current Freelancing	Employee firm size (Quarter 3, size = 0-19 employees) Employee firm age (Quarter 2, age = 2 to 3 years) Privately held firm rate Employee firm age (Quarter 1, age = 0 to 1 years) Employee firm size (Quarter 2, size = 0-19 employees) Employee firm age (Quarter 1, age = 11+ years) Employee establishment age (0 years) State government workers Employee firm age (Quarter 2, age = 0 to 1 years)
State x Gender	Gig Work	No additional covariates
State x Gender	Nascent Entrepreneurship	Business or professional income tax returns Effective sample size Employee establishment age (11 to 15 years) Self-employment tax returns New employer business actualization Employee firm size (Quarter 4, size = 500+ employees) Employee firm size (Quarter 2, size = 500+ employees)
State x Gender	Withdrawn Entrepreneurship	Federal government workers
State x Gender	Former Business Ownership	Employee firm age (Quarter 1, age = 11+ years) Self-employed: incorporated
State x Race	Current Business Ownership	Employee firm size (Quarter 1, size = 0-19 employees) Employee firm age (Quarter 2, age = 6 to 10 years) Employee establishment age (5 years)

State x Race	Former Freelancing	Self-employed: incorporated Unpaid family workers employee firm size (Quarter 1, size = 500+ employees) Self-employed: non-incorporated employee firm age (Quarter 4, age = 0 to 1 years) Employee firm size (Quarter 1, size = 20-49 employees)
State x Race	Non-entrepreneurship	Employee establishment age (26+ years) Employee firm size (Quarter 3, size = 250-499 employees) Employee firm size (Quarter 4, size = 250-499 employees) Employee firm age (Quarter 1, age = 0 to 1 years) Employee firm size (Quarter 4, size = 0-19 employees) Private, not for profit, wage and salary workers employee firm size (Quarter 2, size = 0-19 employees) Employee firm age (Quarter 4, age = 11+ years) Employee establishment age (11 to 15 years) Employee establishment age (21 to 25 years) Employee firm age (Quarter 3, age = 11+ years)
State x Race	Current Freelancing	Employee firm size (Quarter 1, size = 20-49 employees) Employee firm size (Quarter 1, size = 500+ employees) Employee firm size (Quarter 2, size = 0-19 employees) Employee firm size (Quarter 2, size = 20-49 employees) Employee firm size (Quarter 3, size = 500+ employees) Employee firm size (Quarter 4, size = 250-499 employees) Employee firm age (Quarter 2, age = 11+ years) Employee firm age (Quarter 3, age = 11+ years)
State x Race	Gig Work	Employed employee firm size (Quarter 4, size = 250-499 employees) Employee firm size (Quarter 4, size = 500+ employees) Employee firm age (Quarter 4, age = 11+ years)
State x Race	Nascent Entrepreneurship	Employee firm size (Quarter 1, size = 20-49 employees) New employer business actualization Self-employed: not incorporated employee firm size (Quarter 3, size = 20-49 employees) Business or professional income tax returns employee firm size (Quarter 4, size = 500+ employees)

State x Race	Withdrawn Entrepreneurship	Employee firm size (Quarter 2, size = 20-49 employees) Employee firm size (Quarter 2, size = 500+ employees) Employee firm size (Quarter 3, size = 250-499 employees)
State x Race	Former Business Ownership	Self-employed: incorporated
MSA X Race P(R/E)	Current Business Ownership	Employee firm size (Quarter 1, size = 500+ employees)
Msa X Race P(R/E)	Withdrawn Entrepreneurship	Employee firm size (Quarter 1, size = 20-49 employees) State government workers
State x Race	Withdrawn Entrepreneurship	No additional covariates
State x Gender	Withdrawn Entrepreneurship	Employed resident rate Working age population Employee firm size (Quarter 2, size = 20-49 employees)