Gender and Racial Disparities in STEM Workforce: A Machine Learning Approach

Luxin Tian

May 25, 2020

Abstract

While previous studies have elaborated on the persisting gender and racial gaps in STEM workforce engagement, there lacks consensus on the classification standards of STEM occupations. In an effort to clarify the boundary of STEM fields, we construct a feature-based machine learning classifier that learns from rich features of occupations and bottom-up responses from workers that indicate their STEM engagement. The classifier achieves 84% accuracy on the test set and suggests that 7.6% of occupations are mislabelled by the official STEM classification. Using the prediction of the feature-based classifier, we examine the gender and racial gaps in STEM workforce engagement in the US in 2018. While the regression coefficients suggest female and racial minorities are significantly underrepresented in the STEM workforce, we find the estimations based on the official STEM classification are prone to overestimate the disparities. We also find the disparities are heterogeneous across states in the US.

Contents

1 Data, Methods, and Results

1.1 Classifying STEM Occupations

1.1.1 Data

To address the limitation of previous classification standards, we take a feature-based approach that integrates multiple data sources to construct an eclectic classification system for predicting STEM occupations. Concretely, we merge the Occupational Information Network (O*NET) database with the American Life Panel (ALP) Survey data. The merged data matches the bottom-up responses from workers who indicate whether the specific job they are doing should be considered as STEM with a comprehensive set of feature variables that characterize all the occupations coded in SOC categories.

The ALP surveys, conducted by RAND, interview a panel of more than 6,000 nation-wide representatives through the internet each year who are sampled based on probabilistically models (Pollard and Baird, 2017). We mainly select the responses to two surveys, MS480 and MS436, which are available on RNAD's website. In MS480, interviewees were directly asked to report whether they are working in a STEM job with the definition of a STEM job described in the question. In MS480, interviewees were also asked to indicate their occupations coded in the SOC categories. Since the panel representatives continuously participate in the surveys, the records on the individual level can be matched across different surveys in the ALP dataset. The two questions in MS480 and MS436 are described in the following way (Pollard and Baird, 2017):

> MS436 Q3 occupation: In which category is your current occupation?
>
> MS480 C2 Whether or not you have a STEM degree, are you currently working in a STEM job? Note that this relates to the tasks you do, and not the industry you work in. For example, an engineer for a bioengineering research firm would be in a STEM job, but an administrative assistant at the same bioengineering research firm would NOT be in a STEM job.

(1 Yes, I am currently working in a STEM job; 2 No, I am not currently working in a STEM job.)

After merging the responses of these two questions and dropping any records with missing data, we obtained a set of labeled data that consists of 1424 records.

To extend the bottom-up STEM labels to out-of-sample occupations, we match the ALP survey data with the O*NET database. The O*NET contains a comprehensive set of descriptive variables that characterizes all aspects of features of the nearly 1,000 occupations in the US economy. The variables are organized into the O*NET Content Model (Peterson et al., 1999) that constitutes 42 files belonging to six main categories of features in regards to occupations and workers. We selected three files that are relevant to identifying STEM occupations from the database, Knowledge, Skills, and Work Activities, and joined the feature variables on the SOC codes of each occupation. For each feature variable, the original data file stores the value of the feature and a series of descriptive statistics of the samples that are used to quantify the feature. We reshaped the data to a wide panel and group any multiple measures of each feature of each occupation by taking the mean of the values in each group and obtained a matrix of 774 occupations with 112 feature variables in total.

Finally, we match the ALP data with the O*NET to construct a labeled dataset for training a machine learning model. After dropping any rows with missing data, the final dataset contains 1279 records with 109 feature variables, and 24.7% of the responses are labeled as STEM.

1.1.2 Model

We build a pipeline for training a machine learning classifier on the merged labeled data and predicting the STEM labels for each occupation in the O*NET dataset. The pipeline contains three steps: standardization, feature selection, and classification.

First, we standardize the feature variables by removing the mean value of each variable and then dividing the difference by the standard deviation of the variable. This step ensures that all the feature variables are centered around zero and maintain

variance in similar orders. Otherwise, some variables in higher orders of magnitude can dominate the cost function in the supervised learning algorithm and prevent the model from learning other features.

Second, we perform feature selection with Recursive Feature Elimination (RFE) to prevent overfitting. RFE selects feature variables to be included in the feature matrix for further classifications by recursively eliminate feature variables and train an estimator on each of the selected subsets of features.

We use a Support Vector Classifier (SVC) with a linear kernel as the estimator for the REF algorithm. An SVC is a Support Vector Machine classifier that constructs a maximal margin hyperplane to separate the training observations of different classes such that the perpendicular distances from the observations to the separating hyperplane are maximized. Concretely, an SVC estimator is the solution to find the maximal margin M and a hyperplane defined by $\mathbf{x}\beta^T$, where $\mathbf{x}$ is the feature matrix and β is the parameter vector $(\beta_0, \beta_1, \ldots, \beta_p)$ in (1).

$$\begin{aligned} \max_{\beta_0,\beta_1,\ldots,\beta_p,M} \quad & M \\ \text{s.t.} \quad & \sum_{j=1}^{p} \beta_j^2 = 1 \\ & y_i \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) \geq M \quad \forall i = 1, \ldots, n \end{aligned} \tag{1}$$

For each subset of features, the SVC gives a vector β, which is orthogonal to the separating hyperplane. Each element in the vector β specifies the importance of the corresponding feature, based on which the features with the least importance are eliminated from the current subset of features. After recursively repeating the elimination until the prespecified number of features are left in the matrix. We further use 5-fold cross-validation to select the optimal number of features to be included for further classification. The cross-validation result of feature selection is shown in Figure 2. We also tried other feature selection methods and selected RFE with Linear SVC based on its joint performance with the Logistic Regression Classifier over other pipeline designs. The model selection scores are shown in Table 1.

Third, we train a classification model on the merged data with bottom-up re-

sponses, which is then used for predicting the STEM labels of all the unlabelled occupations in the O*NET dataset. We experimented with several classifiers, the Logistic Regression Classifier, the Support Vector Machine Classifier, and tree-based classifiers such as the Decision Tree Classifier and the Random Forest Classifier. We randomly split the merged data into a 0.8 training set and a 0.2 test set. For each experimental classifier, we performed a grid search with 5-fold cross-validation to tune the hyperparameters. The cross-validation scores for each pipeline are shown in Table 1.

Finally, we selected the Logistic Regression Classifier, which achieves mean accuracy scores of 0.8407 and 0.8438 on the training and test sets, respectively. Concretely, a Logistic Regression fits the following model 2: $\forall i \in 1, 2, \ldots, n$:

$$\begin{aligned} \mathrm{Z_i} &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \\ \Pr(\mathrm{STEM}_i = 1) &= \mathrm{sigmoid}(\mathrm{Z_i}) \end{aligned} \tag{2}$$

We train the Logistic Classifier with the optimal parameter from the grid search and predict the STEM labels for all the occupations in the O*NET dataset. With the default 0.5 activation threshold of the sigmoid function, the classifier does not perform well, as it labels far fewer STEM occupations than the SOC official classification. We experimented with different probability thresholds to label the STEM occupations and plot the number of STEM labels against the thresholds value in Figure 3. The slope of the curve indicates the sensitivity of the selection of the threshold value. We select 0.43 as the threshold value where the sensitivity shows a sharp swift, indicating there is can exist a sharp gap between STEM and non-STEM occupations. We obtained a set of 774 occupations associated with predicted probability and labels.

1.1.3 Results

Using the classifier we construct, we can predict the score of each occupation that indicates the extent to which it is STEM-like. As is shown in Figure 1, our feature-based classifier gives a similar result to the SOC official classification. However, 5.5% occupations that are identified as STEM by SOC are not labeled as STEM by

our classifier, and 2.1% occupations that are not approved as STEM are actually showing STEM features instead. These occupations are referred to as periphery STEM occupations by Anderson et al. (2018), which reflects the limitations of the traditional classification standard in identifying STEM occupations that are emerging and fading out.

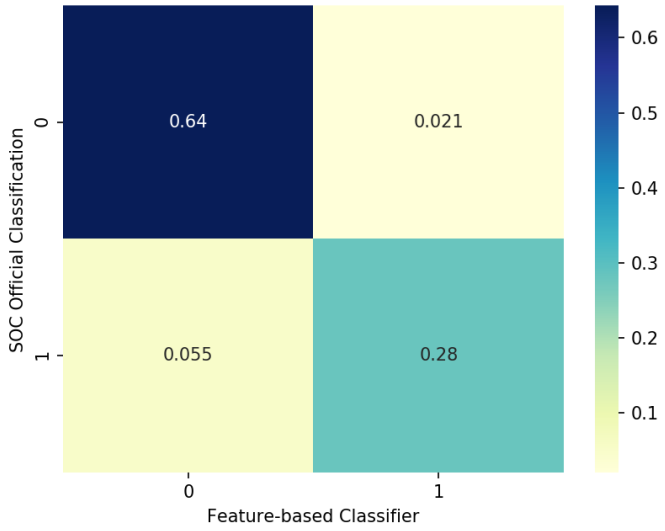


Figure 1: Confusion Matrix of Feature-based Classification and SOC Official Classification

1.2 Measuring Gender and Racial Disparities

Our empirical analysis of the gender and racial disparities in the STEM workforce replicates several previous studies, which adopt either the SOC official classification or the Brookings scoring approach. We demonstrate the potential of our feature-based classification of STEM occupations by evaluating the difference in results given by different classification approaches. Concretely, we fit a linear Logistic Regression model with the American Community Survey (ACS) data by the US Census Bureau.

1.2.1 Data

We use the 2018 one-year data of the ACS, which contains 3,214,539 individuals across the United States, with sample weights specified. We select sex, racial codes, and SOC code variables to match our classification result on the SOC code. After dropping any records with missing data, we obtained a cross-section that consists of 1,247,722 samples. The descriptive statistics is shown in Table 2.

1.2.2 Model

As a demonstration of using our classification results, we fit reduced-form simple linear models to identify the gender and racial disparities in the STEM workforce, without examining the causal effects and mechanisms. Concretely, we specify a Linear Regression model and a Logistic Regression model in (3) and (4).

$$\mathrm{Z} = \alpha + \beta_1 \mathrm{Gender} + \beta_2 \mathrm{Race} + \beta_3 \mathrm{Gender} \times \mathrm{Race} + \gamma \mathrm{X} + \epsilon \tag{3}$$

$$Y = \mathrm{sigmoid}(\mathrm{Z}) \tag{4}$$

Gender is a dummy variable which equals 1 when the individual is male and 0 if female. Race is a matrix of dummy variables indicating the racial groups. X is a matrix of control variables, including age, a binomial term of age, and school attainments. Y is the dependent variable. When Y is a discrete binary variable, such as the predicted STEM label or the official STEM label, we will fit the Logistic Regression that combines (3) and (4), and when Y is the natural log of the predicted probability of being STEM, we fit a weighted OLS regression model (3). We assign the samples with the weights in the ACS data.

1.2.3 Results

We fit the regression models using both the nation-level data and state-level data. Through the nation-level regressions, we identify the gender and racial disparities in STEM workforce across the United States in 2018. The result indicates that females

remain significantly underrepresented in the STEM workforce, using as the dependent variable either the official STEM labels, our feature-based labels, or the natural log probability of being engaged in STEM occupations given by our classifier. The result is robust across model specifications with different sets of covariates. However, the coefficient of the gender variable in the with our feature-based STEM labels is lower than that of the model with the official labels, indicating the traditional classification approach may have resulted in researchers overestimating the gender disparities in STEM workforce engagement.

We also identified that racial minority groups such as African American, American Indian, Alaska Native, American Indian, and Hawaiian or Pacific Islanders are significantly underrepresented in STEM workforce compared to White Americans, and females in these minority groups are further underrepresented in comparison to male. However, Asian people show a significantly higher probability of engaging in STEM workforce than White Americans, but female Asians are still less likely to engage in STEM occupations that male Asians. Again, we find that the results may have been overestimated by researchers using the traditional classification approach. Full nation-level regression results are shown in Table 3, 4, and 5.

Using the state-level data, we only fit the WLS regressions with the predicted probability as a STEM score of each occupation, with all control variables included. The result shows that gender disparities are heterogeneous across the states in the US. We plot the regression coefficients in Figure 4. States with the most significant gender disparities are South Dakota, Vermont, Iowa, and North Dakota.

References

Anderson, Drew, Matthew Baird, and Robert Bozick, "Who Gets Counted as Part of America's STEM Workforce? The Implications of Different Classification Approaches for Understanding the Gender Gap in STEM," 2018.

Peterson, Norman G, Michael D Mumford, Walter C Borman, P Jeanneret, and Edwin A Fleishman, *An occupational information system for the 21st century: The development of O* NET.*, American Psychological Association, 1999.

Pollard, Michael and Matthew D Baird, "The RAND American Life Panel: Technical Description," 2017.

A Tables and Graphs

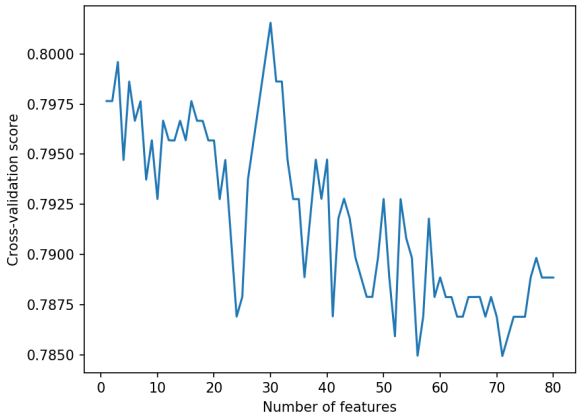


Figure 2: Cross-validation scores of RFECV with Linear SVC

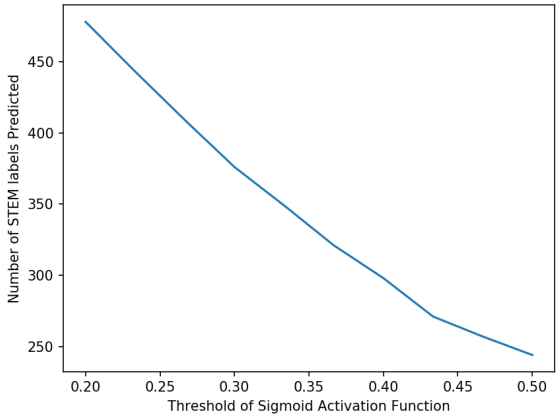


Figure 3: Sensitivity of Number of STEM Labels Predicted on Activation Threshold Value

Table 1: Performance of Experimental Feature Selection Methods and Classifiers

Pipeline	Feature Selection	Classifier	Best CV Score on Training/Test Data	Precision by Labels	Recall by Labels	F1 by Labels
1	None	Random Forest	0.86706, 0.80859	0.88, 0.58	0.86, 0.63	0.87, 0.61
2	LASSO with CV	Random Forest	0.86706, 0.79685	0.87, 0.57	0.87, 0.57	0.87, 0.57
3	Ridge with CV	Random Forest	0.86706, 0.79297	0.87, 0.56	0.86, 0.57	0.86, 0.56
4	Logistic with CV	Random Forest	0.86706, 0.79688	0.87, 0.57	0.87, 0.57	0.87, 0.57
5	RFECV with Linear SVC	Random Forest	0.86706, 0.8125	0.88, 0.63	0.87, 0.57	0.87, 0.60
6	None	Logistic with CV	0.82111, 0.84375	0.86, 0.75	0.95, 0.50	0.90, 0.60
7	RFECV with Linear SVC	Logistic with CV	0.84066, 0.84376	0.87, 0.72	0.93, 0.55	0.90, 0.62

Table 2: Descriptive Statistics of Variables

	Weight	Male	Age	SOC STEM label	Pred Prob	African Amrcan	American Indian	Alaska Native	American Indian or Alaska Native	Asian	Hawaiian/Pacific	Other Races	Two or More Races
count	1247722.00	1247722.00	1247722.00	1247722.00	1247722.00	1247722.00	1247722.00	1247722.00	1247722.00	1247722.00	1247722.00	1247722.00	1247722.00
mean	101.26	0.56	44.70	0.24	0.24	0.08	0.01	0.00	0.00	0.06	0.00	0.03	0.02
std	81.89	0.50	15.88	0.43	0.21	0.28	0.08	0.02	0.04	0.23	0.04	0.18	0.15
min	1.00	0.00	16.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	53.00	0.00	31.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50%	80.00	1.00	45.00	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
75%	122.00	1.00	58.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
max	2097.00	1.00	96.00	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 3: Weighted Logistic 1: Gender and Racial Disparities in STEM Workforce Engagement

	Logistic: Z = SOC Official STEM Labels				
VARIABLES	I	II	III	IV	V
Male	1.0634***	1.0647***	1.0623***	0.9741***	1.0245***
	(0.0046)	(0.0047)	(0.0047)	(0.0051)	(0.0057)
Age		0.1046***	0.1036***	0.0529***	0.0529***
		(0.0009)	(0.0009)	(0.0010)	(0.0010)
Age-squared		-0.0010***	-0.0010***	-0.0005***	-0.0005***
		(0.0000)	(0.0000)	(0.0000)	(0.0000)
African American			-0.1776***	0.0050	-0.0902***
			(0.0081)	(0.0090)	(0.0181)
American Indian			-0.6622***	-0.2725***	-0.0960
			(0.0322)	(0.0358)	(0.0635)
Alaska Native			-1.1766***	-0.5372***	-0.5636
			(0.1617)	(0.1783)	(0.3763)
American Indian or Alaska Native			-0.7308***	-0.2433***	0.1750
			(0.0763)	(0.0850)	(0.1349)
Asian			0.5535***	0.2870***	0.6425***
			(0.0083)	(0.0092)	(0.0148)
Hawaiian/Pacific			-0.7973***	-0.4503***	-0.1780
			(0.0699)	(0.0766)	(0.1292)
Other Races			-1.1044***	-0.4970***	-0.3140***
			(0.0173)	(0.0191)	(0.0332)
Two or More Races			-0.2013***	-0.1577***	0.0665**
			(0.0152)	(0.0167)	(0.0286)
Female African American					0.1233***
					(0.0209)
Female American Indian					-0.2542***
					(0.0766)
Female Alaska Native					0.0285
					(0.4276)
Female American Indian					-0.6428***
					(0.1722)
Female Asian					-0.5575***
					(0.0186)
Female Hawaiian/Pacific					-0.4020**
					(0.1596)
Female of Other Races					-0.2669***
					(0.0403)
Female of Two or More Races					-0.3287***
					(0.0350)
const	-1.7948***	-4.1593***	-4.1092***	-5.0293***	-5.0611***
	(0.0039)	(0.0192)	(0.0193)	(0.0564)	(0.0564)
Controls	N	N	N	Y	Y
Num of observations	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722

Standard errors in parentheses. *p <0.1; **p <0.05; ***p <0.01.

Table 4: Weighted Logistic 2: Gender and Racial Disparities in STEM Workforce Engagement

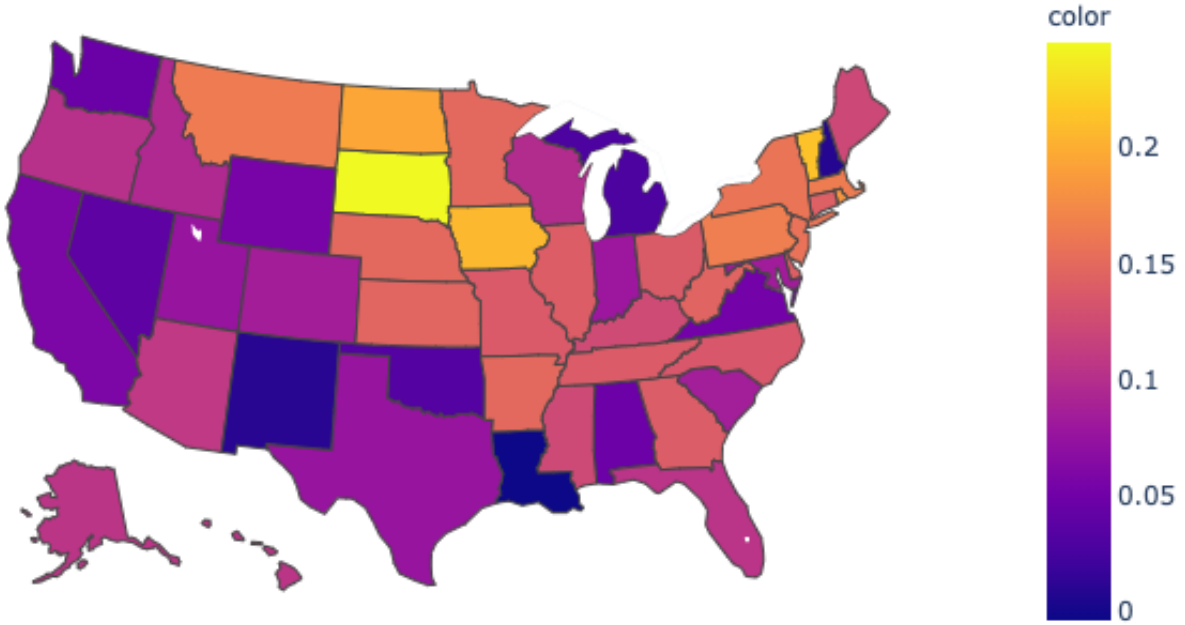
	Logistic: Z = Featured-based Predicted Label				
VARIABLES	I	II	III	IV	V
Male	0.9616***	0.9604***	0.9578***	0.8433***	0.8896***
	(0.0048)	(0.0048)	(0.0048)	(0.0053)	(0.0059)
Age		0.1028***	0.1018***	0.0474***	0.0474***
		(0.0009)	(0.0009)	(0.0011)	(0.0011)
Age-squared		-0.0010***	-0.0010***	-0.0004***	-0.0004***
		(0.0000)	(0.0000)	(0.0000)	(0.0000)
African American			-0.2552***	-0.0759***	-0.1722***
			(0.0085)	(0.0095)	(0.0192)
American Indian			-0.7168***	-0.3044***	-0.0560
			(0.0343)	(0.0382)	(0.0659)
Alaska Native			-1.0376***	-0.3001*	0.0914
			(0.1615)	(0.1815)	(0.3216)
American Indian or Alaska Native			-0.6945***	-0.1662*	0.2721**
			(0.0789)	(0.0887)	(0.1378)
Asian			0.5710***	0.2894***	0.5885***
			(0.0084)	(0.0093)	(0.0151)
Hawaiian/Pacific			-0.7825***	-0.4007***	-0.0575
			(0.0730)	(0.0803)	(0.1301)
Other Races			-1.1592***	-0.5305***	-0.3212***
			(0.0187)	(0.0206)	(0.0349)
Two or More Races			-0.2006***	-0.1502***	0.0881***
			(0.0157)	(0.0173)	(0.0294)
Female African American					0.1240***
					(0.0222)
Female American Indian					-0.3620***
					(0.0806)
Female Alaska Native					-0.5457
					(0.3871)
Female American Indian					-0.6896***
					(0.1784)
Female Asian					-0.4674***
					(0.0189)
Female Hawaiian/Pacific					-0.5218***
					(0.1644)
Female of Other Races					-0.3093***
					(0.0429)
Female of Two or More Races					-0.3529***
					(0.0362)
const	-1.8634***	-4.2024***	-4.1476***	-5.0016***	-5.0329***
	(0.0040)	(0.0199)	(0.0200)	(0.0605)	(0.0605)
Controls	N	N	N	Y	Y
Num of observations	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722

Standard errors in parentheses. *p <0.1; **p <0.05; ***p <0.01.

Table 5: WLS: Gender and Racial Disparities in STEM Workforce Engagement

	log(Pr{STEM=1})				
VARIABLES	I	II	III	IV	V
Nale	0.2399***	0.2323***	0.2263***	0.1151***	0.1075***
	(0.0016)	(0.0016)	(0.0016)	(0.0014)	(0.0016)
Age		0.0499***	0.0497***	0.0258***	0.0257***
		(0.0003)	(0.0003)	(0.0003)	(0.0003)
Age-squared		-0.0005***	-0.0005***	-0.0002***	-0.0002***
		(0.0000)	(0.0000)	(0.0000)	(0.0000)
African American			-0.1299***	-0.0636***	-0.1522***
			(0.0025)	(0.0022)	(0.0034)
American Indian			-0.2356***	-0.0811***	-0.0326**
			(0.0105)	(0.0092)	(0.0133)
Alaska Native			-0.3062***	-0.0608	-0.0153
			(0.0492)	(0.0433)	(0.0603)
American Indian or Alaska Native			-0.3250***	-0.1392***	-0.0929***
			(0.0215)	(0.0190)	(0.0270)
Asian			0.1850***	0.0234***	0.0396***
			(0.0033)	(0.0029)	(0.0045)
Hawaiian/Pacific			-0.2702***	-0.1256***	-0.1001***
			(0.0190)	(0.0167)	(0.0247)
Other Races			-0.4396***	-0.1604***	-0.0964***
			(0.0038)	(0.0035)	(0.0047)
Two or More Races			-0.0775***	-0.0556***	-0.0470***
			(0.0049)	(0.0044)	(0.0065)
Female African American					0.1513***
					(0.0045)
Female American Indian					-0.0934***
					(0.0184)
Female Alaska Native					-0.0961
					(0.0866)
Female American Indian					-0.0917**
					(0.0379)
Female Asian					-0.0265***
					(0.0059)
Female Hawaiian/Pacific					-0.0474
					(0.0336)
Female of Other Races					-0.1372***
					(0.0068)
Female of Two or More Races					-0.0158*
					(0.0087)
const	-1.9956***	-3.1434***	-3.0968***	-3.1684***	-3.1644***
	(0.0012)	(0.0061)	(0.0061)	(0.0096)	(0.0096)
Controls	N	N	N	Y	Y
R-squared	0.0175	0.0501	0.0653	0.2757	0.2767
Num of observations	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722

Standard errors in parentheses. *p <0.1; **p <0.05; ***p <0.01.



The values are the coefficients of the sate-level WLS regression with the natural log of the predicted probability of being engaged in STEM as the dependent variable. States with the most significant gender disparities are South Dakota, Vermont, Iowa, and North Dakota.

Figure 4: State-level Gender Disparities in STEM Workforce Engagement in the US (2018)