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

Research Proposal

Luxin Tian

April 14, 2020

Research Question

- How does the presence of people from different gender and racial groups distribute in STEM and non-STEM workforce?
- How does the difference in earnings between people in STEM and non-STEM workforce vary among different gender and racial groups?
- How do these differences change over years?

Existing Research

- Beede D N, Julian T A, Langdon D, et al. Women in STEM: A gender gap to innovation[J]. Economics and Statistics Administration Issue Brief, 2011 (04-11).
- Carnevale A P, Smith N, Melton M. STEM: Science Technology Engineering Mathematics. State-Level Analysis[J]. Georgetown University Center on Education and the Workforce, 2011.
- Landivar L C. Disparities in STEM employment by sex, race, and Hispanic origin[J]. Education Review, 2013, 29(6): 911-922.
- …

- Greater shares of male than women work in STEM workforce.
- Racial minorities are underrepresented in STEM workforce.
- Female has lower retention in STEM workforce.
- Wage promotion of STEM occupations is greater for women.
- …

How is the boundary of STEM workforce defined?

Previous research: usually uses top-down classifications

- the official categorization by the federal government

Caveats: potential measurement errors due to inaccurate classifications can lead to substantial estimation bias (Anderson et al, 2018)

- Not all workers in a designated STEM occupation actually do STEM jobs
- Not all workers in a designated non-STEM occupation actually don't do STEM jobs

How is the boundary of STEM workforce defined?

Previous research: Top-down classifications

- the official categorization by the federal government

A new `ALP` survey (2017): provides bottom-up responsess

- self-reported classifications

Anderson et al (2018):

- Different classifications lead to tremendously different empirical results regarding gender gaps between STEM and non-STEM field.

STEM measurement: A Machine-learning Classifier

Available data source:

- the offical classification standard
- a bottom-up survey to workers (`RAND Aamerican Life Panel`)
- characteristic data of all occupations (`O*Net`)

		workers' response	
		STEM	non-STEM
official classifications	STEM	(A) certain STEM occupations	(B) periphery occupations
	non-STEM	(C) periphery occupations	(D) certain non-STEM occupations

- Use (A)(D) as the training set
- Use `O*Net` as the features

STEM measurement: A Machine-learning Classifier

Available data source:

- the offical classification standard
- a bottom-up survey to workers (`RAND Aamerican Life Panel`)
- characteristic data of all occupations (`O*Net`)

		workers' response	
		STEM	non-STEM
official classifications	STEM	(A) certain STEM occupations	(B) periphery occupations
	non-STEM	(C) periphery occupations	(D) certain non-STEM occupations

- Use (A)(D) as the training set
- Use `O*Net` as the features

STEM measurement: A Machine-learning Classifier

Machine learning algorithms

- Logistic classifier
- K-nearest Neighbors classifier
- Support Vector Machines classifier
- Naive Bayes classifer
- Tree-based classifers

Construct the STEM indicator for further regression analysis:

- predicted outcome: $\texttt{STEM} \in \{0, 1\}$
- probability estimation as STEM score: $\texttt{STEMScore} \in [0, 1]$

Identification Strategy: Regression Analysis

Question 1: STEM job engagement by gender and race.

- Descriptive statistics: difference between groups
- Logistic regression:

$$\mathrm{Z} = \alpha + \beta_1 \mathrm{Gender} + \beta_2 \mathrm{Race} + \beta_3 \mathrm{Gender} \times \mathrm{Race} + \gamma \mathrm{X} + \epsilon$$
$$\Pr(\mathrm{STEM} = 1) = \mathrm{sigmoid}(\mathrm{Z})$$

- Tree-based regression

Data

- US Census Bureau: `Current Population Survey (CPS)`

Identification Strategy: Regression Analysis

Question 2: heterogeneous earning gaps by gender and race.

- Linear regression:

$$\begin{aligned}\text{Wage} = \alpha + &\beta_1 \text{Gender} + \beta_2 \text{Race} + \beta_3 \text{STEMScore} + \\ &\beta_4 \text{Gender} \times \text{Race} + \\ &\beta_5 \text{Gender} \times \text{STEMScore} + \\ &\beta_6 \text{Race} \times \text{STEMScore} + \\ &\beta_7 \text{Gender} \times \text{Race} \times \text{STEMScore} + \\ &\gamma \text{X} + \epsilon\end{aligned}$$

Hypotheses and Expectations

Expectations

- Machine-learning classifers provide more accurate classifications of STEM occupations
- Using the predicted indicators leads to more reliable empirical results regarding disparities between STEM and non-STEM workforce.

Hypothesis

- Previous research may have **overestimated/underestimated** the gender and racial gaps between STEM and non-STEM workforce.