Methods and Initial Results

Yaling Xu

May 2019

1 Data

I collect the data from five year American Community Survey (ACS) from 2015 to 2017. (link: https://usa.ipums.org/usa-action).

2 Data Cleaning and Restriction Process

2.1 Occupation Data:

I choose OCC1990 (OCC1990 is a modified version of the 1990 Census Bureau occupational classification scheme) as my occupation basis in order to keep a consistent classification of occupations using the 1990 coding scheme as its starting point.

And then I make some occupation recoding in oder to keep consistent for these there years. For example, I recode actuaries as mathematicians. More details about the recoding shows in the Appendix 1.

I used the way that Dorn (2009) developed to construct 301 consistently defined occupations. I dropped the occupations that are not consistent over time, and then compiled them to 17 occupations: Managers; Natural, Math and Computing Science Professionals; Architects and Engineers; Health Professionals; Educators; Business Professionals (Accountants, HR, etc); Legal Professionals; Social Scientists; Writers and Artists; Technicians; Religious/social workers; Office clerks; Market salespersons; Precision Production, Operators, craft, repair Professionals; Agricultural workers; Not in

LF or in Military. But I won't consider the Not in LF or in Military occupation in the following analysis since the sample is too small and limited.

2.2 Major Data:

Here, to properly assign the major that is most predictive of an individual's employment when they have double-majored, I assign the highest earnings major as their primary major. There are several majors where the first and second major have the same earnings maximum. For these, I call the major as the respondent that listed as first.

2.3 Other Restrictions:

I limit the data to native-born Americans who have bachelor degree and age 25 to 57. I drop the self-employed individuals. I only keep individuals whose usual work time is at least 30 hours or 3 weeks last year. I also discount the wage to the 2018 level using CPI index. Then I dropped people with missing field for first major.

3 Map Variables to Theoretical Constructs

I define overwork as working over 50+ hours. And I construct the degree of overwork in an occupation as the share of overwork among males. As discussed by Goldin, workplace flexility is a multi-dimensional concept the include working hours per week, flexibility of leaving, interaction with colleagues and other factors. In our analysis, I focus on the overwork (work for long hours) since it is the most significant factor affecting workplace flexibility (Goldin, 2004).

The key variable in my analysis is the occupational distribution of females given specific demographic group, and it is measured as the share of females working in an occupation given demographic group in each decade.

Demographic groups are characterized by age, marital status, and fertility variables since these variables could be regarded as cost of committing working long hours.

4 Descriptive Statistics

There are significant variations in overwork across occupations. The average overwork percentage is 23.9 percent. One one hand, the top three occupations that require longest working hours are Legal Professionals, Precision Production, Operators, craft, repair Professionals and Managers. They all have overwork percentage that over 30 percent. On the other hand, the percentage of overwork among Religious/social workers, Office Clerks and Health Professionals are lower than 15.5 percent.

Legal Professionals, precision production, operators, craft, repair professionals and managers still appear to make up the largest proportion of overworking occupation among females. In other words, for females working in these three occupations, they have over 29 percent probability to work for over 50 hours per week. The results are same for males. Religious/social workers, office clerks and health professionals still have lowest probability to overwork for females, while natural, math and computing science Professionals seem easier for males than religious/social workers.

5 Model

For main analysis, I plan to examine the relationship between changes in the prevalence of overwork and the female occupational distribution based on following baseline regression:

$$\frac{female\_gr_{it}}{female\_gr_{t}} = \alpha + \beta * share\_male\_overworkk_{it} + \gamma * \frac{control\_gr_{it}}{control\_gr_{t}} + \delta * X_{it} + \pi_i + \pi_t + \epsilon_{ict} \tag{1}$$

where i refers to an occupation and t refers to each decade. $female_{gr}$ is the female demographic group of interest; $control_{gr}$ means another demographic group in the same age range but facing relatively lower costs of working long hours. For example, female age 25-44 without children may have lower cost than female who is at the same age but have children controlling other factors.

6 Initial Empirical Analysis

In this section, I test the equation (1) for females given a demographic group that married, with at least one child, with at least bachelor degree and age from 15 to 40. This demographic group is assumed as the group that has the largest cost for long hours working commitment.

For initial analysis, I control for occupation and year fixed effects. Then I add the log earnings as the control variable manually. For the further analysis, I will use feature selection to include more control variables.

Table 1 reports the results for initial analysis. Column (1) reports the coefficient of β only controlling for year and occupation fixed effects. In column (2), log earnings is added as another control variable. These two coefficients are both significant in 10 percent level, which means that as the share of male overworking in an occupation increases, the employment share of females with bachelor degree and at least one child in the same occupation decreases.

However, our results may capture some occupation-specific demand shocks that are correlated with the overwork pattern and female occupational distribution. In order to address this concern, I include controls for other demographic groups: who are likely to be less sensitive to the overwork: share of individuals of the same age working in occupation i. Here, I control for males, single females and married females without children in column (3), (4) and (5). The coefficients are essentially unchanged.

On one hand, since the control group should face the similar demand shocks with our interest group, our result after controlling control group should reflect the lower willingness of married women with children remain in one occupation that requires long hours commitment to some extent. On the other hand, it helps us to provide assurance that the overwork in an occupation is not merely proxying for other skills valued by the occupation, for which there might be gender differences, such as competitiveness and risk-taking. In sum , this result in accordance with our previous assumption: once overwork prevail in an occupation, well-educated females who have children are likely to "flee" from the occupation.

7 Computational Methods

The variable selection is important for the high-dimensional datasets. In these situation it is hard to easily say which of the variables are relevant and which ones are irrelevant, and on the other hand, it is difficult, due to dimensionality issues, to build and interpret a model that takes into consideration all the variables. For these reasons the feature selection is an important task since it selects the features by ranking them on how useful they are for the model, to compute the usefulness score statistical test and correlation results are used.

In this paper, I plan to use LASSO do feature selection my control variables. However, I met a problem here and I hope to ask for professor's help in the office hour.

Since the dependent variable I care about is the occupational distribution of college-educated female with children and age from 25 to 57. I only have 16 occupations here, which means that dependent variable only appears 16 times. However, the potential control variables data is in individual level. I don't know how to do feature selection on the imbalanced dataset.

Figure 1: Magnitude of Overwork per Occupation

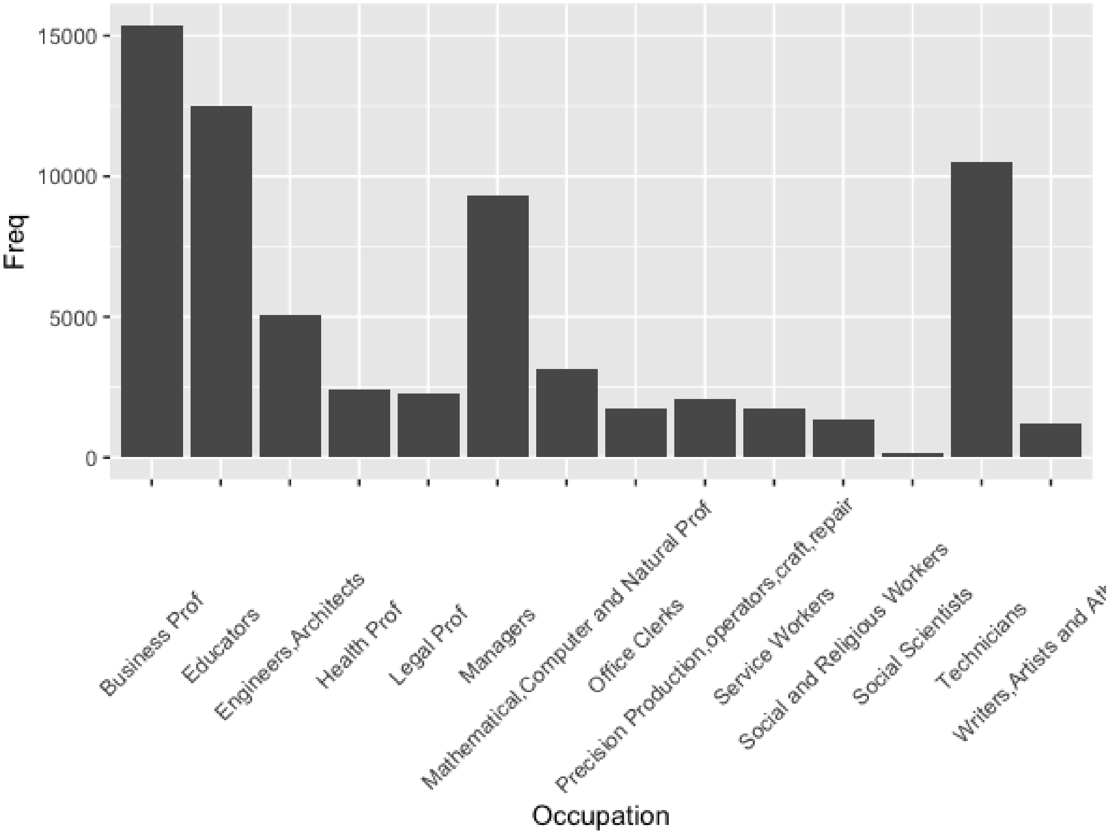


Figure 2: Magnitude of Overwork per Occupation(1=male, 2=female)

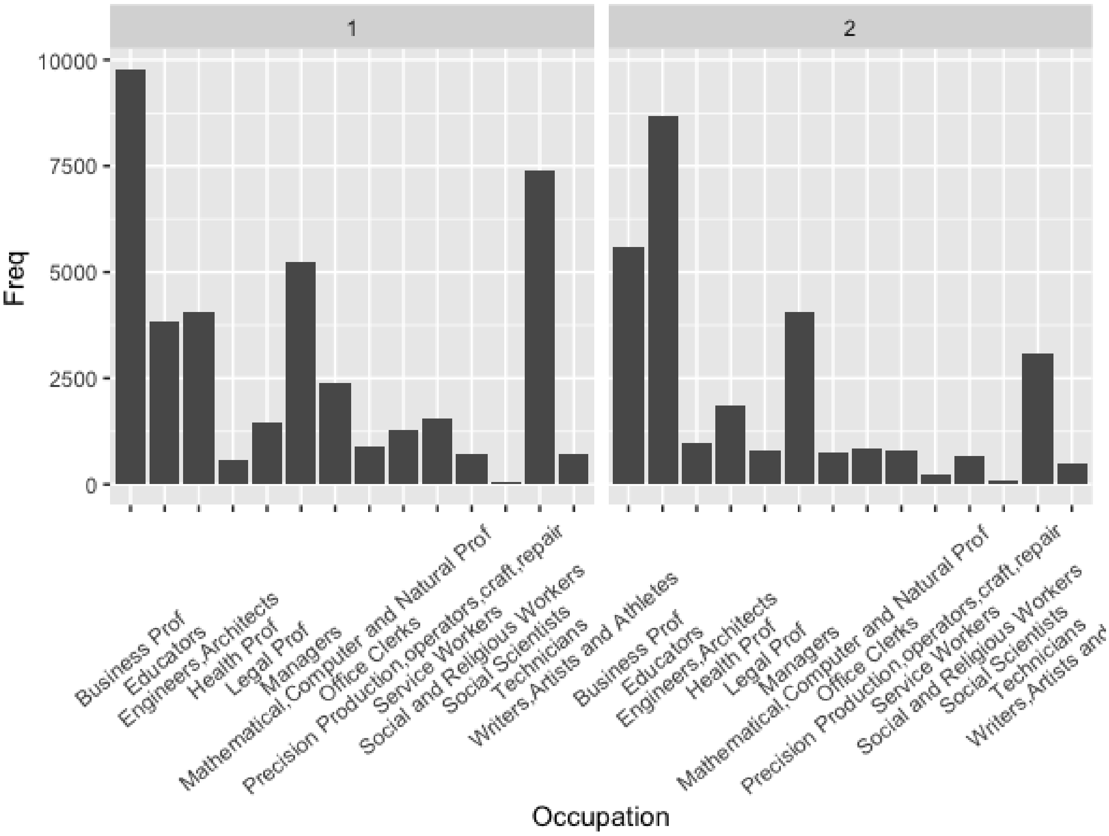
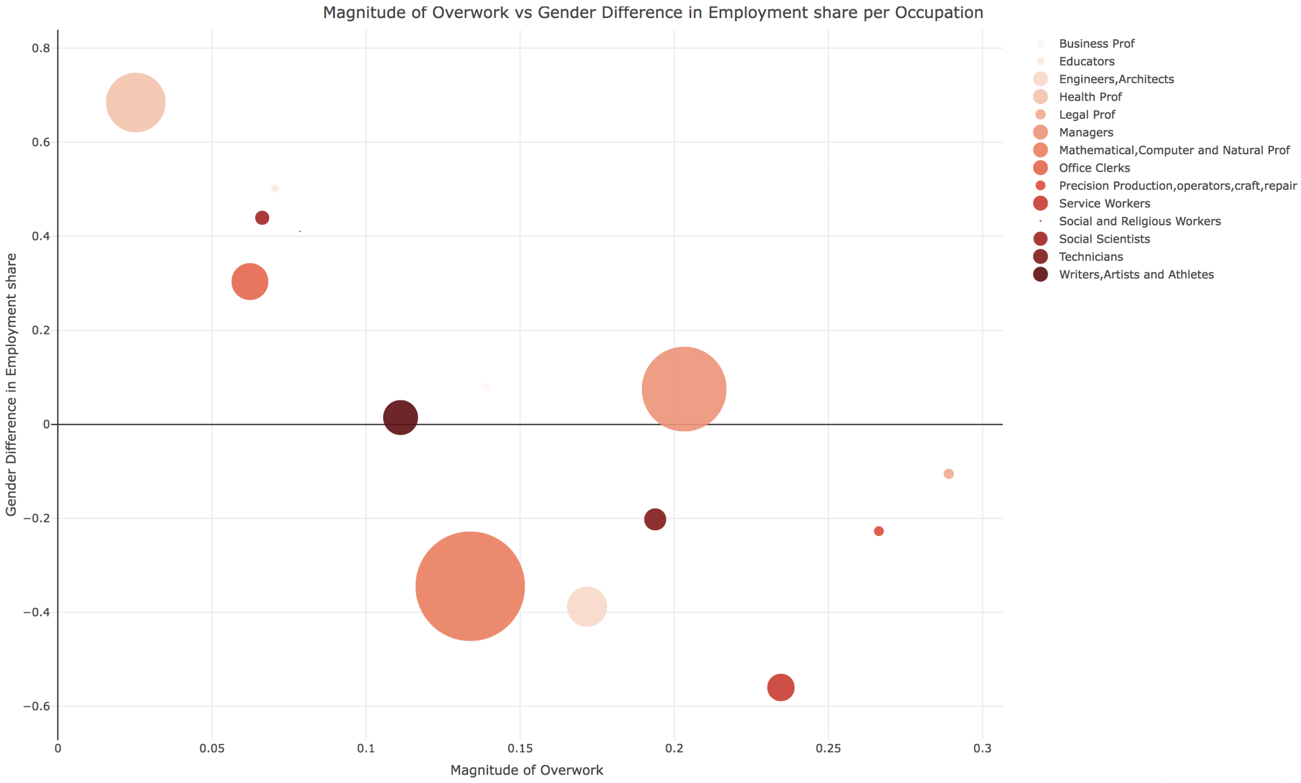


Figure 3: Magnitude of Overwork vs Gender Difference in Employment share per Occupation



Occupation	share of male	share of female	share of overwork	share of male in overwork	share of female in overwork
Business Prof	0.460099	0.539901	0.2185882	0.6363518	0.3636482
Educators	0.2491277	0.7508723	0.229625	0.3066219	0.6933781
Engineers,Architects	0.6939343	0.3060657	0.2127211	0.8071287	0.1928713
Health Prof	0.1576793	0.8423207	0.1077594	0.2345169	0.7654831
Legal Prof	0.5526839	0.4473161	0.4500994	0.6422261	0.3577739
Managers	0.4625329	0.5374671	0.3597306	0.5648806	0.4351194
Mathematical,Computer and Natural Prof	0.6722128	0.3277872	0.1763441	0.758665	0.241335
Office Clerks	0.3481325	0.6518675	0.121494	0.512761	0.487239
Precision Production, operators,craft,repair	0.6136082	0.3863918	0.4313402	0.6175908	0.3824092
Service Workers	0.7798913	0.2201087	0.2672101	0.8779661	0.1220339
Social and Religious Workers	0.2948041	0.7051959	0.151255	0.5196507	0.4803493
Social Scientists	0.280303	0.719697	0.1590909	0.4166667	0.5833333
Technicians	0.60114	0.39886	0.2748666	0.7050989	0.2949011
Writers,Artists and Athletes	0.4927764	0.5072236	0.1909548	0.5822368	0.4177632

Table 1: Descriptive statistics of overwork and occupation distribution

	Married Female sage 25-57 With Children				
Share of males age 25 to 57 working 50+ hours	-0.186*	-0.158**	0.154**	-0.121**	-0.149**
Share of Males of the same age working in occupation i			0.839***		
Share of Single Females of the same age working in occupation i				0.732***	
Share of Married Females without children of the same age working in occupation i					0.577***
Standard deviation of log male and female wages		X	X	X	X
Occupation FE	X	X	X	X	X
Year FE	X	X	X	X	X
R-Squared	0.960	0.971	0.934	0.912	0.943

Table 2: Relationship between the magnitude of overwork and well-educated female occupational choices