Research Proposal

Luning Li

2021/3/29

Introduction

With the development of the feminist movement in the past decade, more and more women choose to enter the workplace. However, the gender pay gap has attracted wide attention. In 2017, women working full time and year-round in the United States typically were paid just 80 percent of what men were paid, a gap of 20 percent (Fontenot et al., 2018). The cause of gender pay gap is controversal. This kind of dispute mainly focuses on whether this gap is due to gender discrimination, or to the difference objective factors between male and female. The purpose of this project is to first investigate to what extend gender factor can affect the wage, and then provide women with advice to get a high-paying job by establishing a regression model on wage income. It may take another decade for women to get equal pay for equal job as men; however, they could contribute more to other independent variables of wage to get themselves a higher wage.

Focus on people at the age of 25-35 in Ohio, who worked above 40 weeks in 2018. We try to find the subjective factors that influence wage. Bayesian hierarchy model is introduced on county level means of wages in Ohio. The advantage of a hierarchy model is that it provides a flexible way to explain the county level means. The results from hierarchy model pulls the estimates of county level means towards the population mean and therefore, less sensitive to noises compared to no pooling method.

Data

The data set used in this research is the American Community Survey 2019 Sample extracted from IPUMS USA (www.ipums.org). It involves the following variables:

- `AGE`: Age of the respondents. Filter the young respondents (age 25-35).
- `SEX`: Sex of the respondents. 1 if male, 2 if female.
- `WKSWORK1`: The number of weeks that respondents worked last year. Filter respondents with `WKSWORK1`$>$ 40.
- `UHRWORK1`: The number of hours per week that the respondent usually worked, if the person worked during the previous year. Filter respondents with `UHRWORK1`>0.
- `INCWAGE`: The total pre-tax wage and salary income - that is, money received as an employee - for the previous year with 999999 = N/A and 999998 = Missing. Filter respondents with `INCWAGE`>0.
- `OCCSCORE`: A constructed variable that assigns occupational income scores to each occupation.
- `PWSTATE2`: The state in which the respondent's primary workplace was located. Filter the respondents working in Ohio (code 39).
- `PWCOUNTY`: The county (or county equivalent) where the respondent worked. 0 if not available. Filter the

The independent variable that we are interested in is the average weekly salary (on log scale) that one respondents received last year. That is:

- $Y = \log\left(\frac{\text{INCWAGE}}{\text{WKSWORK1}}\right)$.

The relevant dependent variables are:

- `SEX`: gender of the respondents;
- `OCCSCORE`: occupational income score;
- `UHRSWORK`: usual hours worked per week.

In addition, a hierarchical structure will be applied on the county level mean. i.e. assume that the county level means of the weekly salary are indepedent samples coming from the same distribution.

Data Preprocessing:

When extracting the data from IPUMS, choose the case where age of respondents is between 25 and 35, and the primary working State to be Ohio. Then preprocess the data for further analysis:

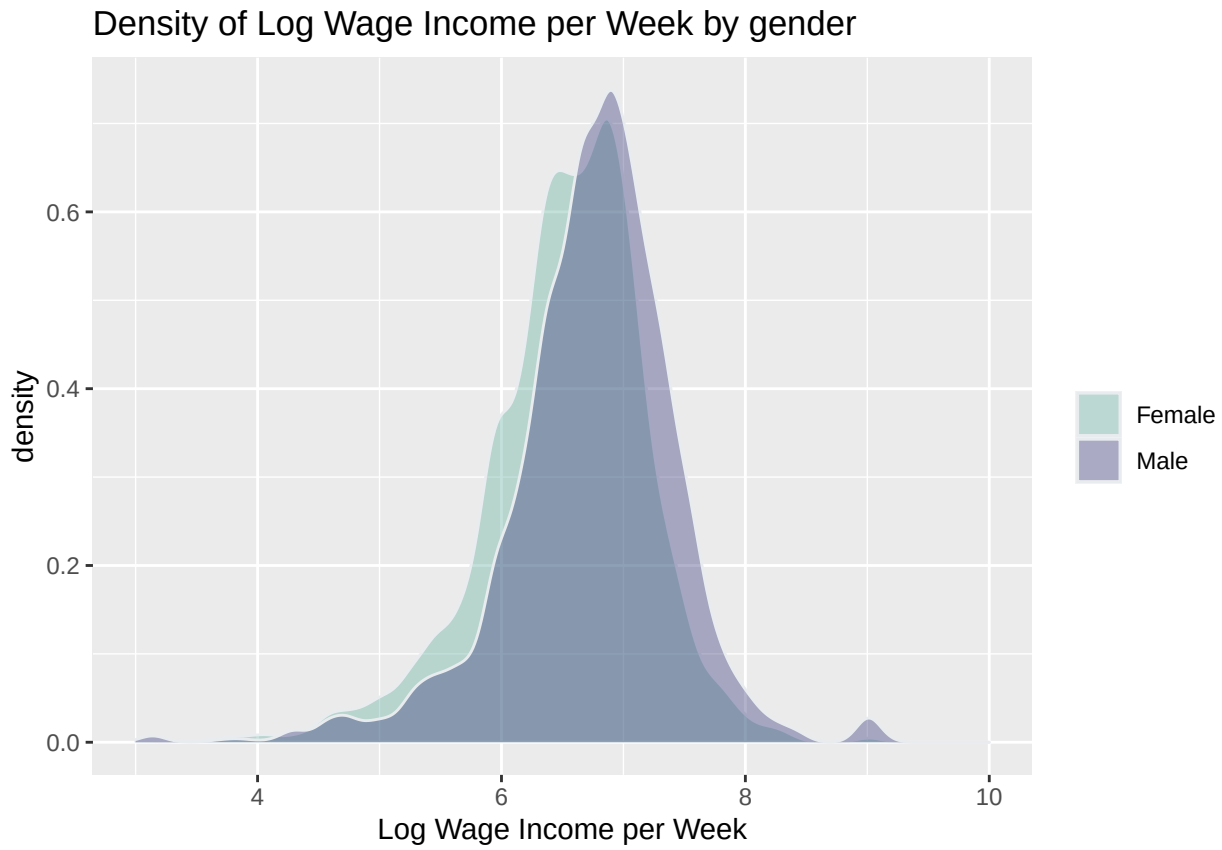
- Select the respondents whose weeks worked last year (`WKSWORK`) is greater than 40.
- Select the respondents whose average hours worked per week (`UHRSWORK`) is greater than 0.
- Select the respondents whose wage and salary income (`INCWAGE`) is greater than 0.
- Select the respondents whose occupational income score (`OCCSCORE`) is greater than 0.

After checking, missing values exist for variable `PWCOUNTY`. Those respondents worked in county not identifiable from public-use data was removed from our research. We only focused on the employees working within an identifiable county.

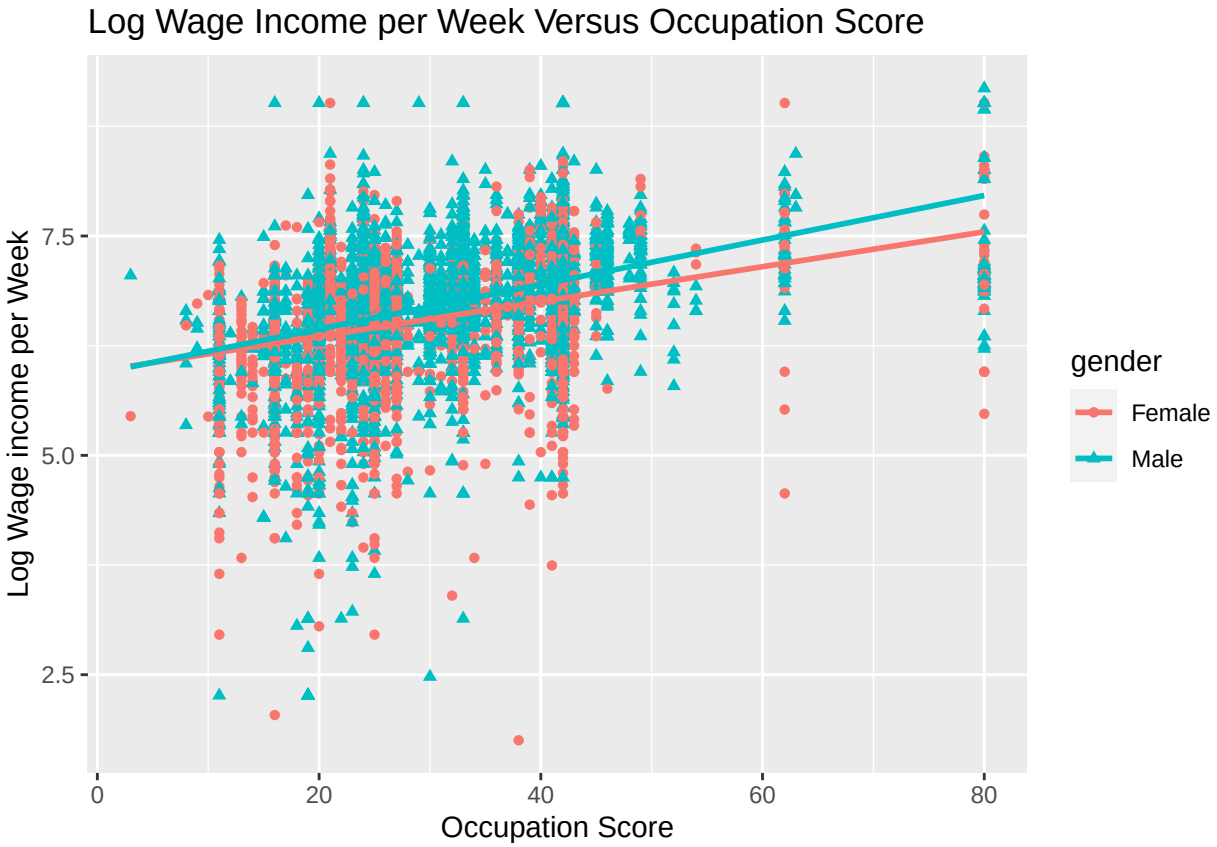
Then the data frame used to fit the model looks like:

gender	UHRSWORK	OCCSCORE	COUNTY	Y
Female	40	25	15	6.483
Female	28	42	15	5.825
Male	30	18	15	5.559
Female	40	23	15	5.847
Female	48	23	15	6.453
Male	40	23	15	6.763

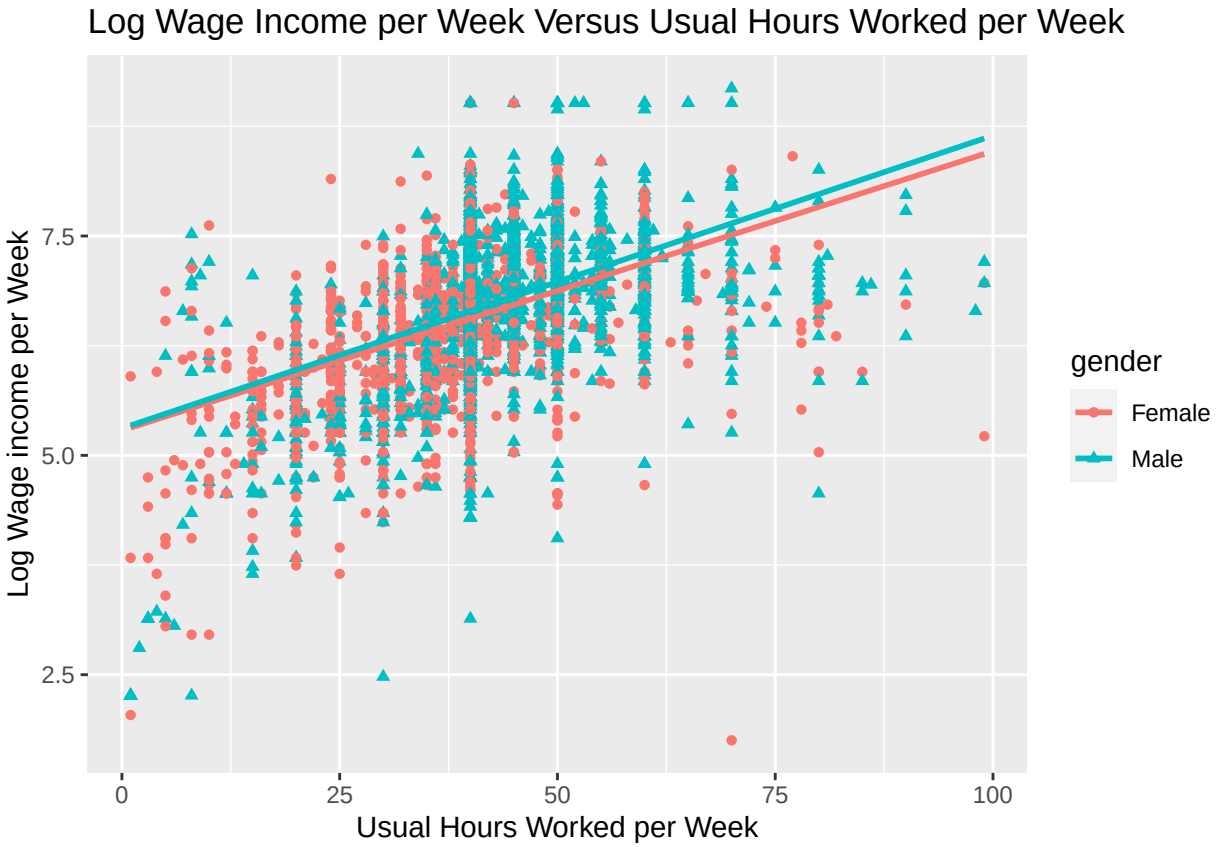
Exploratory Data Analysis



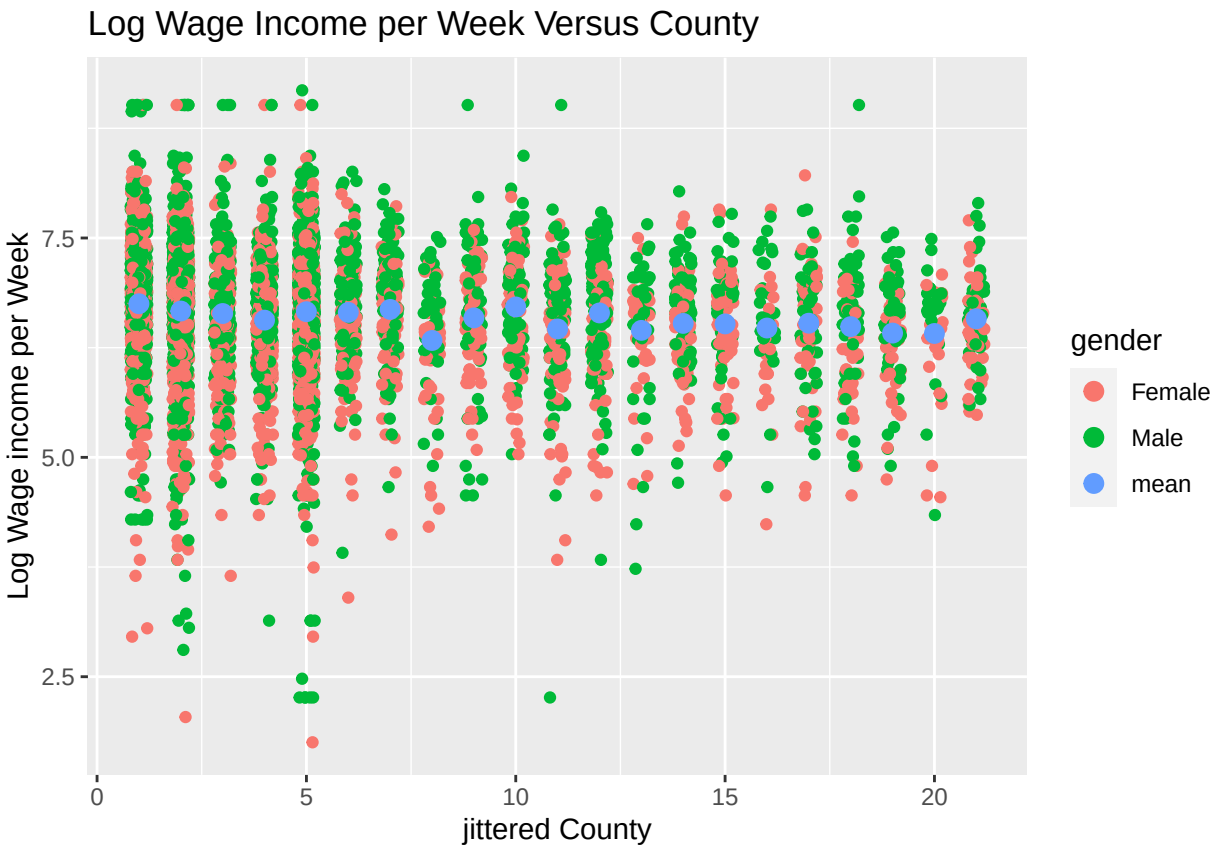
The density plot of log weekly wage income (`Y`) illustrates the gender pay gap exists in Ohio. This density from female is more left-tailed than that from male, and therefore, females are generally paid less than the males. Therefore, this research will treat `gender` as a predictor for wage.



The scatter plot of `Y` against the occupation score (`OCCSCORE`) shows a positive correlation. Therefore, `OCCSCORE` is taken as a predictor for the wage. Moreover, the solid lines represent the linear regression lines between those two variables from each gender. The two regression lines are close to each other, so no interaction term between `OCCSCORE` and `gender` is considered in the model.



The scatter plot of `Y` against the usual hours worked per week (`UHRSWORK`) shows a positive correlation. Therefore, `UHRSWORK` is taken as a predictor for the wage. Again, the solid lines represent the linear regression lines between those two variables from each gender. The two regression lines are close to each other, so no interaction term between `UHRSWORK` and `gender` is considered in the model.



The county means of observed `Y` are distributed around 6.5. It is reasonable to assume that these means come from the same distribution. This justifies our hierarchical structure on county level mean of `Y`. Note that in this plot, the County code is jittered to make the observations from each county more visible.

Reference

Fontenot, K., Semega, J., & Kollar, M. (2018). Income and Poverty in the United States: 2017. Washington: U.S. Census Bureau.

Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D010.V11.0