Analysising the Gender Pay Gap In The U.S Workplace

Erdong Zhang

2020/12/20

Analysising the Gender Pay Gap In The U.S Workplace

Erdong Zhang

2020/12/22

Github repo link:

https://github.com/ErdongZ/Sta304---Final-Project.git

Abstract

The gender pay discrepancy is a social problem that is shown as a gender imbalance in the workplace's payments. This study applies the approach of propensity score matching and suits a linear regression model to find out the difference of gender income in the U.S. workplace. The result is significant and reveals that holding age, race, state of residence, education attainment, hours of work, and worker fixed class. The man, on average, continues to receive a 37.3% higher annual pre-tax income than the female. The outcome shows that gender wage inequality is sexism against women in the workplace and the implications of the disparity in different ways impact the U.S. economy and the lives of women.

Keywords

Gender pay gap, propensity score matching, Discrimination, Disparity

Introduction

The U.S used to be a pioneer in tackling gender income inequality. Of the rich countries, it was the first country to pass laws restricting occupational gender inequality. However, it currently has a more significant gender pay gap than most OECD countries (Glynn, 2018). Additionally, Glynn (2018) notes that the prevalence of wage disparity between men and women would directly and implicitly impact the U.S economy. In addition, she suggests that the gender wage gap decreases family wealth, weakens spending power, and directly impacts single-parent families with low and medium incomes. Moreover, Farrell and Glynn (2013) identify many factors. For instance, differences in work experiences, occupations, and industries could explain a portion of the wage disparity. They stress, though that nearly half of the reasons related to gender wage inequality are not measurable, and may be overt bias or unconscious prejudice against women that prohibits them from bargaining in the workplace for a better payment. Hence, drawing a causal inference in this context is the key to uncover the unobservables and discover the impact of gender discrimination on wage disparities in the U.S workplace. One way to determine the causal inference is the propensity score matching method. Propensity score matching is a popular, powerful, and robust method for determining the causal effect on observational data (Arbour et al., 2014). This report adopts the propensity score matching method. It complies with the latest U.S microdata that aims to justify that gender discrimination is pervasive in the U.S current work environment and causes vast income disparity between men and women. Since the last century, the U.S government has implemented laws and regulations to address the gender pay gap issue and improve the corresponding legislation. However, the pay gap has not been significantly narrowed so far. According to Bleiweis (2020), "the gender pay gap has only closed by 4 cents for every $1 earned by men in more than a decade." Also, she points out that absent any reforms in current legislations, closing the gender pay gap is infeasible in the next forty years. Thus, by precisely specify the causal effect between gender discrimination and the gender pay gap, this report also suggests that the U.S government should take further measures to eliminate gender discrimination and protect females' equality in the workplace. The following Methodology section contains Data and Model subsections. The Data section provides information regarding data collection and a baseline characteristic table of the data. The model section shows the specific equations of the model and discusses relevant features that enter the model. The result of the propensity score analysis is provided in the Result section. Interpretations of the result and further discussions related to the gender pay gap are in the Discussion section.

Methodology

Data

This report uses the 2019 American Community Survey (ACS) dataset retrieved from Integrated Public Use Microdata Series (IPUMS-USA). IPUMS-USA provides easily accessible U.S census microdata that consists of abundant demographic and economic variables. According to the ACS information guide (United States Census Bureau, 2017), the data is collected through the following approaches: The United States Census Bureau randomly selects household addresses across the country with each has 1/480 probability of been determined, and the same household should not be chosen more than once every five years. The majority of the data is collected via mailing the letters to the randomly selected addresses that invite people living in the address to participate in an online survey. If the Census Bureau does not receive a completed response in a few weeks, the Census Bureau will send an additional paper questionnaire to the address. Besides, the Census Bureau will conduct a personal interview survey for people living in group housing such as nursing homes, prisons, and college dormitories. To deal with non-respondents or uncompleted surveys, the Census Bureau (1) sends a field representative to conduct a personal interview with the address that did not respond to the survey online or through the mail. (2) conducts a telephone follow-up. The data used in this report contains 31437 randomly selected observations from the 2019 ACS dataset. It has eight variables, which are age, sex, race, educational attainment, state of residence, class of worker, weeks of working last year, and annual pre-tax personal income (will be referred to as personal income for simplicity), where sex, race, educational attainment, state of residence, and a class of worker are categorical variables; age, weeks of working, and personal income are numerical variables. A new variable called treatment is created to determine the propensity score to perform the propensity score matching method to achieve the propensity score matching method. The variable treatment is a dummy variable with a value of 1 if the sex is female and equal to 0 if the sex is male. Table 1 provides a baseline characteristics table of the data. It shows the mean and standard deviations of the variables, which aims to provide a general overview of the data.

Model

For the purpose of testing the hypothesis that there is a gender pay gap in the U.S. workplace, this report incorporates a logistic regression model and multiple linear regression model in computing the propensity score and ascertaining the gender effect on wages, respectively. Both models are run in the R-studio, more specifically, The logistic regression model to calculate the propensity score reads as:

$$log(\frac{p}{1-p})_i$$
$$= \tilde{\beta}_0 + \tilde{\beta}_1 age_i + \tilde{\beta}_2 wkswork1_i + \tilde{\beta}_3 educd_i + \tilde{\beta}_4 stateicp_i + \tilde{\beta}_5 classwkrd_i$$
$$+ \tilde{\beta}_6 race_i + \epsilon_i$$

where p represents the probability of being assigned to the treatment group for each I in the sample. The estimated coefficient $\tilde{\beta}_1$ captures the change in log odds for one year increasing in age and $\tilde{\beta}_2$ represents the change in the log odds for each additional week of working. Besides, $\tilde{\beta}_3$ to $\tilde{\beta}_6$ represent the effect of a specific category on changes in log odds. Lastly, ϵ is the error term. Since the dependent variable, treatment is a binary dummy variable. The logistic regression model will provide a more precise estimate for a binary dependent variable than other models such as linear regression models. It is the main reason for choosing the logistic regression model in determining the propensity score. After finishing calculating the propensity score for each individual in the sample and the pair matching process, this report proposes a multiple linear regression model to estimate the gender income disparity. The model has the following format:

$$log(inctot_i)$$
$$= \beta_0 + \beta_1 age_i + \beta_2 race_i + \beta_3 educd_i + \beta_4 stateicp_i + \beta_5 classwkrd_i + \beta_6 sex_i + u_i$$

where β_1 represents the change in pre-tax annual personal income for the individual i in the sample as getting one year older. β_2 to β_5 specifies the magnitude effect of demographic and geographic factors on the pre-tax annual personal income for the individual i. β_6 is the primary interest of the study in this report, which estimates the gender pay gap, and u_i is the error term of the linear regression. The model is well predicted and fits the data since it has an R-squared value of 0.529, and model diagnostic plots (See Figure 1) do not violate linear assumptions. Significantly, residuals are randomly spread out around a horizontal line without a distinct pattern, as shown in the residuals vs. fitted plot. Besides, the normal Q-Q plot indicates the assumption of residual normality is reasonable. Additionally, in the scale-location plot, dots are randomly equal spread out along the asymptotic horizontal line with a few outliers, suggesting the homoscedasticity of residual. Lastly, all the points inside the Cook's distance, including outliers in the residual vs. leverage plot, there is no influential sample that affects the model prediction. Overall, the multiple linear regression model is adequate for estimating and feasible determining the gender pay gap. Moreover, the report chooses to use the logarithm of personal income (log(inctot)) as the dependent variable instead of personal income. The log income's estimated coefficients can provide a more explicitly fraction form result to show that females tend to earn β_6 cents less for every $1 made by a male on average, holding age, race, educational attainment, state of residence, and a class of worker fixed.

Result

Statistical results of the multivariate analysis is presented below.

```
## # A tibble: 92 x 5
##    term                                estimate std.error statistic
  p.value
##    <chr>                                  <dbl>     <dbl>     <dbl>
    <dbl>
##  1 (Intercept)                           6.99    0.0791      88.4
0
##  2 age                                   0.0235  0.000335    70.1
0
##  3 raceblack/african american/negro      0.0342  0.0562       0.609
0.542
##  4 racechinese                           0.156   0.0681       2.28
0.0224
##  5 racejapanese                          0.0928  0.105        0.886
0.376
##  6 raceother asian or pacific islander   0.121   0.0588       2.05
0.0403
##  7 raceother race, nec                   0.0928  0.0597       1.56
0.120
##  8 racethree or more major races        -0.0262  0.105       -0.248
0.804
```

```
##  9 racetwo major races                      0.162   0.0631       2.57
0.0102
## 10 racewhite                                0.215   0.0537       4.00
0.0000634
## # ... with 82 more rows
```

The model's estimate implies that assuming age, race, educational attainment, state of residence, class of worker, and weeks of working are equal, females on average tend to earn 37.3% less pre-tax annual income than males in the U.S workplace. The difference is statistically significant in a 1% significance level and also economically significant as females, in general, earn 37.3 cents less for every 1 dollar made by males under Ceteris paribus. The result is based on the propensity score matching method that applies the logistic regression to determine the propensity score and uses multivariate analysis based on the new sample that has been processed by the propensity score matching method. Besides, the estimate β_6 has a p-value less than 0.01 implies the data provides sufficient pieces of evidence to conclude the gender pay gap does exist in the U.S. workplace

Discussion

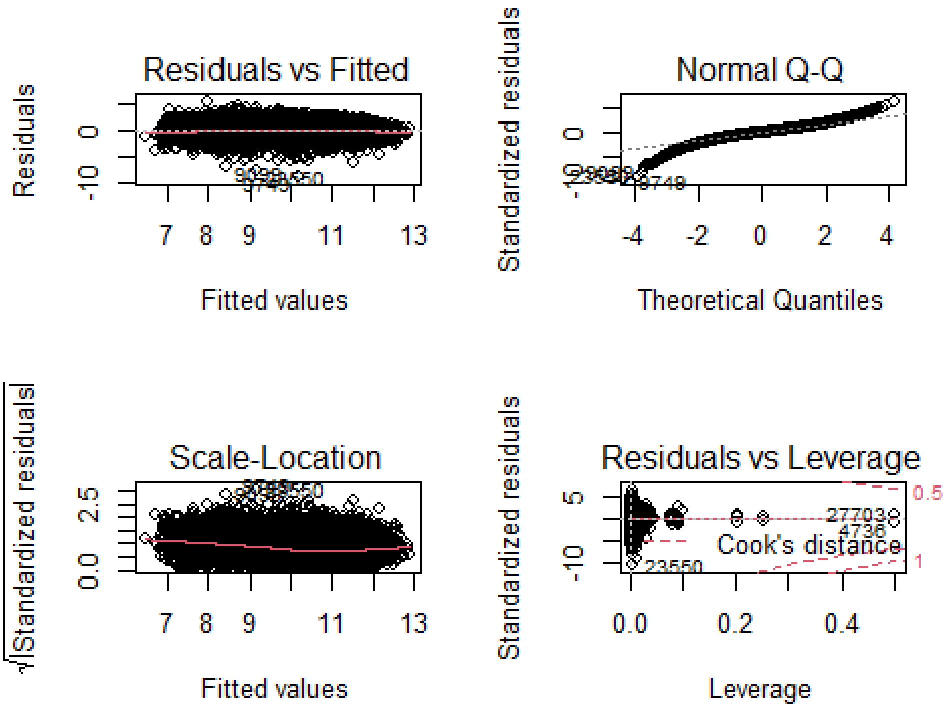
The gender pay gap is a long-standing phenomenon that pervades the workplace in the United States, and since the last century, the appeal for wage equality has not ceased. Many factors cause the gender pay gap; this report interests the gender discrimination effect on income. Moreover, being able to derive a causal inference is essential for an analytical study. To establish the causal association between gender and income inequality, this study adopts the propensity score matching process. Also, the data used in this report provide a significant advantage in deriving the causality as households are randomly selected by the Census bureau. Lastly, the information uses logistic regression to determine the propensity score and propose a multiple linear regression model to estimate the gender pay gap's magnitude. According to the propensity score analysis, holding age, race, education attainment, class of worker, and weeks of working the same, females on average earn 37.3% less than males in the U.S. workplace. It is a huge income gap in practice and has a vital statistical significance. In other words, the result justifies that overt sexism causes females to earn 37.3 cents less for every $1 made by males. Therefore while legislation has been adopted by the U.S. government to reduce the gender wage differential created by gender inequality against women, the statistical outcome suggests that the U.S. government should take more steps to mitigate the impact of misogyny on wages. There may however be concerns of homogeneity in the sample that could theoretically influence the organization and the accuracy of calculating the gender wage difference. Due to privacy issues or other purposes, because much of the data is gathered from online surveys, participants could prefer to distort their personal details, especially income. Increasing the sample size would reduce bias due to misleading responses. Future studies could consider including interaction terms or analyzing the race income disparity.

Appendix

Table 1: Baseline Characteristics Table

```
## 
##                                                       Overall      

##   n                                                    31436       

##   age (mean (SD))                                      43.53 (15.55)

##   sex = male (%)                                       16540 (52.6) 

##   race (%)                                                          

##      american indian or alaska native                    272 ( 0.9) 

##      black/african american/negro                       2698 ( 8.6) 

##      chinese                                             447 ( 1.4) 

##      japanese                                             94 ( 0.3) 

##      other asian or pacific islander                    1371 ( 4.4) 

##      other race, nec                                    1205 ( 3.8) 

##      three or more major races                            91 ( 0.3) 

##      two major races                                     732 ( 2.3) 

##      white                                             24526 (78.0) 

##   educd (%)                                                         

##      1 or more years of college credit, no degree       4731 (15.0) 

##      12th grade, no diploma                              522 ( 1.7) 

##      associate's degree, type not specified             2877 ( 9.2) 

##      bachelor's degree                                  7024 (22.3) 

##      doctoral degree                                     515 ( 1.6) 

##      ged or alternative credential                      1081 ( 3.4) 

##      grade 1                                              12 ( 0.0) 
##      grade 10                                                469 ( 1.5)
##      grade 11                                                620 ( 2.0)
##      grade 2                                                  23 ( 0.1)
##      grade 3                                                  37 ( 0.1)
##      grade 4                                                  26 ( 0.1)
##      grade 5                                                  43 ( 0.1)
##      grade 6                                                 151 ( 0.5)
##      grade 7                                                  40 ( 0.1)
##      grade 8                                                 196 ( 0.6)
##      grade 9                                                 249 ( 0.8)
##      kindergarten                                              4 ( 0.0)
##      master's degree                                        3140 (10.0)
##      no schooling completed                                  283 ( 0.9)
##      nursery school, preschool                                 5 ( 0.0)
##      professional degree beyond a bachelor's degree          742 ( 2.4)
##      regular high school diploma                            6406 (20.4)
##      some college, but less than 1 year                     2240 ( 7.1)
##   stateicp (%)
##      alabama                                                 425 ( 1.4)
##      alaska                                                   67 ( 0.2)
##      arizona                                                 633 ( 2.0)
##      arkansas                                                280 ( 0.9)
##      california                                             3736 (11.9)
##      colorado                                                602 ( 1.9)
##      connecticut                                         363 ( 1.2)
##      delaware                                             96 ( 0.3)
##      district of columbia                                 91 ( 0.3)
##      florida                                            1851 ( 5.9)
##      georgia                                             996 ( 3.2)
##      hawaii                                              137 ( 0.4)
##      idaho                                               150 ( 0.5)
##      illinois                                           1272 ( 4.0)
##      indiana                                             691 ( 2.2)
##      iowa                                                302 ( 1.0)
##      kansas                                              322 ( 1.0)
##      kentucky                                            394 ( 1.3)
##      louisiana                                           399 ( 1.3)
##      maine                                               140 ( 0.4)
##      maryland                                            630 ( 2.0)
##      massachusetts                                       807 ( 2.6)
##      michigan                                            965 ( 3.1)
##      minnesota                                           568 ( 1.8)
##      mississippi                                         279 ( 0.9)
##      missouri                                            591 ( 1.9)
##      montana                                              98 ( 0.3)
##      nebraska                                            204 ( 0.6)
##      nevada                                              292 ( 0.9)
##      new hampshire                                       145 ( 0.5)
##      new jersey                                          819 ( 2.6)
##      new mexico                                          177 ( 0.6)
##      new york                                           1889 ( 6.0)
##      north carolina                                      994 ( 3.2)
##      north dakota                                         88 ( 0.3)
##      ohio                                               1175 ( 3.7)
##      oklahoma                                            331 ( 1.1)
##      oregon                                              403 ( 1.3)
##      pennsylvania                                       1313 ( 4.2)
##      rhode island                                         99 ( 0.3)
##      south carolina                                      457 ( 1.5)
##      south dakota                                        103 ( 0.3)
##      tennessee                                           645 ( 2.1)
##      texas                                              2619 ( 8.3)
##      utah                                                297 ( 0.9)
##      vermont                                              73 ( 0.2)
##      virginia                                            884 ( 2.8)
##      washington                                          774 ( 2.5)
##      west virginia                                       156 ( 0.5)
##      wisconsin                                           564 ( 1.8)
##      wyoming                                              50 ( 0.2)
##   classwkrd (%)
##      federal govt employee                               966 ( 3.1)
##      local govt employee                                2437 ( 7.8)
##      self-employed, incorporated                        1231 ( 3.9)
##      self-employed, not incorporated                    1982 ( 6.3)
##      state govt employee                                1360 ( 4.3)
##      unpaid family worker                                106 ( 0.3)
##      wage/salary at non-profit                          2824 ( 9.0)
##      wage/salary, private                              20530 (65.3)
##   wkswork1 (mean (SD))                                 45.86 (13.41)
##   inctot (mean (SD))                                59372.58 (76355.
24)
##   treatment (mean (SD))                                 0.47 (0.50)
```

Figure 1: Model Diagnostic plots



Reference

Arbour, D., Marazopoulou, K., Garant, D., & Jensen, D. (n.d.). Propensity Score Matching for Causal Inference with Relational Data. [PDF]. University of

Massachusetts Amherst. https://staff.fnwi.uva.nl/j.m.mooij/uai2014-causality-workshop/papers/paper5.pdf.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Bleiweis, R. (2020, March 24). Quick Facts About the Gender Wage Gap.https://www.americanprogress.org/issues/women/reports/2020/03/24/482141/quick-facts-gender-wage-gap/.

Farrell, J., & Glynn, S. J. (2014, July 8). What Causes the Gender Wage Gap? Center for American Progress. https://www.americanprogress.org/issues/economy/news/2013/04/09/59658/what-causes-the-gender-wage-gap/.

Glynn, S. J. (2019, September 25). Gender wage inequality. Equitable Growth. https://equitablegrowth.org/research-paper/gender-wage-inequality/?longform=true.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M. (2020). IPUMS USA: Version 10.0 [Data set]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V10.0

Understanding diagnostic plots for linear regression analysis | university of virginia library research data services + sciences. Retrieved from https://data.library.virginia.edu/diagnostic-plots/

United States Census Bureau (2017, October). American Community Survey Information Guide. [PDF]. U.S Department of Commerce.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Yoshida, K.,& Bartel, A., (2020). tableone: Create 'Table 1' to Describe Baseline Characteristics with or without Propensity Score Weights. R package version 0.12.0. https://CRAN.R-project.org/package=tableone