Occupation in Wage Gap Between Sex

Jiahui Cai*
Department of Mathematics
UC San Diego
j7cai@ucsd.edu

Weiyue Li*
Halıcıoğlu Data Science Institute
UC San Diego
wel019@ucsd.edu

Abstract

Despite the wage gap between sex having been narrowed in recent years, women still only earned 83 cents to every dollar earned by men in 2020 .United States Census Bureau [2021]. On the other hand, a related article claims that there is some occupational segregation in each state in the U.S.Wisniewski [2022]; therefore, in this project, we perform a regression analysis to the IPUMS CPS DatasetRuggles et al. [2020]. In particular, we divide occupations into male-dominated, female-dominated, and equal-dominated and perform a regression analysis to aim for drawing a causal inference of occupation on the wage gap between sex during the COVID and post-COVID era.

1 Introduction

The gender wage gap has persisted over a significant period of time. In 2020, the prevalence of the COVID-19 pandemic resulted in a significant spike in unemployment rates, leading to job loss for many individuals. Under this turbulent economic climate, it is also shocking to see that women are even earning much more than ever compared to men. According to a recent study, women in some major cities are even able to earn more than men after the pandemic Cahn [2022]. Given this new discovery, we are eager to investigate the post-COVID situation in the gender wage gap through our research project. Specifically, we aim to explore the role of occupation in determining the earnings disparity between men and women during the COVID-19 period, spanning from 2020 onwards. After constructing our model, we discovered that the gender wage gap varies significantly across different occupations. Specifically, we found that being a male worker might cause him earn 14.182 percent

[1] Equal contribution to this project.

more than a female worker in female-dominated occupations, 20.099 percent more in male-dominated occupations, and 29.106 percent more in equal-dominated occupations.

2 Discussion of relevant economic theory

A study conducted in the early 2000s revealed that discrimination is a significant factor contributing to the gender wage gap, particularly in relation to the phenomenon of occupational feminization, wherein the representation of women in certain occupations increases over time Asaf Levanon and Allison [2009]. The authors observed that as women enter previously male-dominated occupations, pay tends to decrease relative to other comparable occupations that continue to be male-dominated. This study highlights the importance of understanding the impact of occupational factors on gender pay disparities, which motivated us to focus on the role of occupation in our research. To achieve this goal, we developed a framework that classifies occupations into three categories: male-dominated (characterized by a male representation of 70% or greater), female-dominated (characterized by a male representation of less than 30%), and equal-dominated (characterized by a male representation between 30% and 70%). This classification scheme enables us to examine the relationship between occupational gender segregation and wage differentials, thereby contributing to a more comprehensive understanding of the underlying causes of the gender pay gap.

3 Data description

In this project, we utilized IPUMS CPS Ruggles et al. [2020] dataset to perform our analysis. We restricted the date of the data to only after 2020-01-01 because we are only interested in the COVID and post-COVID periods of the wage gap. Table 1 shows the summary statistics of the original dataset while all the columns are encoded to numerical values despite whether they should be categorical data or not.

Table 1: Summary Statistics of the Original Dataset

	count	mean	std	min	50%	max
YEAR	2977215.0	2.021409e+03	5.897718e-01	2020.0	2021.0	2022.0
SEX	2977215.0	1.512143e+00	4.998526e-01	1.0	2.0	2.0
REGION	2977215.0	2.932945e+01	1.026635e+01	11.0	31.0	42.0
WKSTAT	2977215.0	5.921235e+01	4.173886e+01	11.0	99.0	99.0
EDUC	2977215.0	7.126316e+01	3.978196e+01	1.0	73.0	125.0
OCC	2977215.0	1.983476e+03	2.767045e+03	0.0	0.0	9840.0
INCWAGE	474234.0	2.094580e+07	4.065778e+07	0.0	33000.0	99999999.0

Here is a description of how we cleaned each column in the original dataset:

- Query the YEAR column to make sure the entries left are all greater than or equal to 2020.
- Drop the unlabeled SEX so that there are only males or females in our analysis.

- Drop the nun-labeled REGION and categorize the rest or REGION into northeast, midwest, south, and west.
- Only keeps the full-time worker according to WKSTAT to minimize the noise of part-time salaries, which are generally much less than full-time salaries.
- Categorize EDUC into HS (high school), COLLEGE, and GRAD (graduate) levels.
- Classify Occupations into male-dominant (more than or equal to 70% male), female-dominant (less than 30% male), and equal-dominant (more than to 30% or less than 70% male).
- Take the natural log of INCWAGE, so our regression shows percentage change in the dependent variable given on unit change of independent variables.

Notice that SEX, REGION, EDUC, and OCC are all categorical variables after cleaning, we one-hot-encoded all of them and dropped the first for each column to prevent the multicollinearity issue. We then introduce the interaction terms of Male_In_OCC_Male and Male_In_OCC_Female because we believe the effect of gender on wage changes in various subgroups of the occupation. Table 2 shows the cleaned data for our regression analysis. Figure 1 shows the distribution of income for males vs females, which indicates the wage gap does exist in our cleaned data.

Table 2: Summary Statistics of the Cleaned Dataset

	count	mean	std	min	50%	max
LN_INCWAGE	165721.0	10.829990	0.846155	0.693147	10.839581	14.557447
REGION_NORTHEAST	165721.0	0.153052	0.360039	0.000000	0.000000	1.000000
REGION_SOUTH	165721.0	0.369784	0.482748	0.000000	0.000000	1.000000
REGION_WEST	165721.0	0.281944	0.449947	0.000000	0.000000	1.000000
EDUC_GRAD	165721.0	0.168675	0.374466	0.000000	0.000000	1.000000
EDUC_HS	165721.0	0.308887	0.462036	0.000000	0.000000	1.000000
OCCU_FEMALE_DOMINANT	165721.0	0.240923	0.427645	0.000000	0.000000	1.000000
OCCU_MALE_DOMINANT	165721.0	0.344030	0.475052	0.000000	0.000000	1.000000
MALE	165721.0	0.549104	0.497584	0.000000	1.000000	1.000000
Male_In_OCC_Male	165721.0	0.295310	0.456183	0.000000	0.000000	1.000000
Male_In_OCC_Female	165721.0	0.039536	0.194868	0.000000	0.000000	1.000000

4 Empirical results

Prior to conducting the regression analysis, we computed the empirical average wage gap between genders across the three occupation categories that we have defined. The results are presented in Table 3. Notably, we observe that the wage gap in male-dominated occupations is comparatively smaller than that in female-dominated and equal-dominated occupations, respectively.

This observation underscores the need to examine the dynamics of occupational gender segregation and their effects on pay differentials, as such trends may contribute to the persistence of wage gaps between genders.

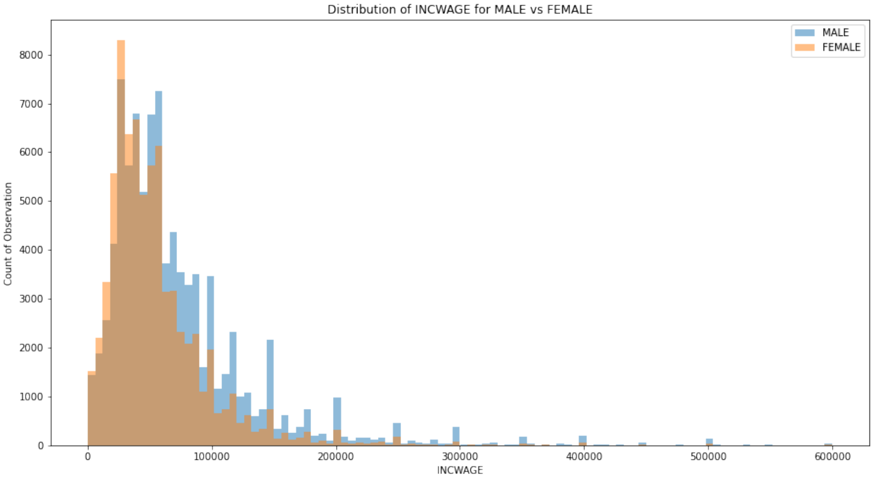

Figure 1: Distribution of INCWAGE for MALE vs FEMALE

Table 3: Wage gap by occupational dominance

Occupation	Wage Gap (Female - Male)
FEMALE_DOMINANT	-14241.637
MALE_DOMINANT	-1158.126
EQUAL_DOMINANT	-23311.763

While there exist numerous factors that may contribute to wage disparities between genders, our study concentrates specifically on the impact of occupation. Accordingly, our regression model includes occupation-related dummy variables and other essential control variables that are relevant to our research question. Equation 1 and 2 present the naive (without interaction terms) and final (with interaction terms) regression model employed in our analysis respectively, and Table 4 displays the resulting output. Notice that all coefficients are statistically significant at 1% significance level according to the t-statistics presented in Table 4

By limiting the focus of our study to occupation-related factors, we aim to isolate the influence of occupational gender segregation on wage gaps. This approach allows for a more precise understanding of the underlying dynamics of gender pay disparities and may provide valuable insights for policy makers in addressing these issues.

$$\begin{aligned}\ln(\mathit{INCWAGE}) = \beta_0 &+ \beta_1 \cdot \mathit{REGION\_NORTHEAST} + \beta_2 \cdot \mathit{REGION\_SOUTH} + \beta_3 \cdot \mathit{REGION\_WEST} \\ &+ \beta_4 \cdot \mathit{MALE} + \beta_5 \cdot \mathit{OCC\_FEMALE} + \beta_6 \cdot \mathit{OCC\_MALE} + \beta_7 \cdot \mathit{EDUC\_GRAD} \\ &+ \beta_8 \cdot \mathit{EDUC\_HS} \qquad (1)\end{aligned}$$

$$
\begin{aligned}
\ln \textit{(INCWAGE)} = {} & \beta_0 + \beta_1 \cdot \textit{REGION\_NORTHEAST} + \beta_2 \cdot \textit{REGION\_SOUTH} + \beta_3 \cdot \textit{REGION\_WEST} \\
& + \beta_4 \cdot \textit{MALE} + \beta_5 \cdot \textit{OCC\_FEMALE} + \beta_6 \cdot \textit{OCC\_MALE} + \beta_7 \cdot \textit{EDUC\_GRAD} \\
& + \beta_8 \cdot \textit{EDUC\_HS} + \beta_9 \cdot \textit{MALE\_IN\_OCC\_MALE} + \beta_{10} \cdot \textit{MALE\_IN\_OCC\_FEMALE}
\end{aligned} \tag{2}
$$

Table 4: Regression table

	(1)	(2)
	LN_INCWAGE	LN_INCWAGE
REGION_NORTHEAST	0.0913	0.0916
	(14.17)	(14.22)
REGION_SOUTH	-0.0369	-0.0365
	(-6.99)	(-6.93)
REGION_WEST	0.0260	0.0268
	(4.69)	(4.82)
MALE	0.216	0.255
	(48.35)	(43.65)
OCC_FEMALE	-0.201	-0.167
	(-39.62)	(-28.11)
OCC_MALE	0.0657	0.114
	(14.08)	(12.00)
EDUC_GRAD	0.518	0.518
	(97.48)	(97.54)
EDUC_HS	-0.456	-0.456
	(-105.29)	(-105.19)
MALE_IN_OCC_MALE		-0.0723
		(-6.62)
MALE_IN_OCC_FEMALE		-0.123
		(-10.32)
Constant	10.78	10.76
	(1924.44)	(1809.98)
Observations	165721	165721

t statistics in parentheses

Table 5 presents the combinations of our target independent variables and their respective interaction terms. This approach aims to enhance our understanding of the dummy variables used in our analysis and to provide insights into how their effects may vary across different groups and conditions.

According to Table 5, we can make the following comparisons between male and female wages:

Table 5: Dummy Table

MALE	OCC_FEMALE	OCC_MALE	Expression	Description	Label
1	1	0	$\beta_4 + \beta_5 + \beta_{10}$	Male in female-dominant occupation	1
1	0	1	$\beta_4 + \beta_6 + \beta_9$	Male in male-dominant occupation	2
1	0	0	β_4	Male in equal-dominant occupation	3
0	1	0	β_5	Female in female-dominant occupation	4
0	0	1	β_6	Female in male-dominant occupation	5
0	0	0	0	Female in equal-dominant occupation	6

- To calculate the wage difference between males and females in female-dominate occupations, we subtract expression 4 from 1: male in female-dominate occupations - female in female-dominate occupations = $(\beta_4 + \beta_5 + \beta_{10}) - \beta_5 = 0.1326229$
- To calculate the wage difference between males and females in male-dominated occupations, we subtract expression 5 from 2: male in male-dominate occupations - female in male-dominate occupations = $(\beta_4 + \beta_6 + \beta_9) - \beta_6 = 0.183144$
- To calculate the wage difference between males and females in equal-dominate occupations, we subtract expression 6 from 3: male in equal-dominate occupations - female in equal-dominate occupations = $(\beta_4) - 0 = 0.2554662$

Based on our model, we get the following conclusions:

- Male earn 14.182 $((e^{0.1326229} - 1) * 100)$ percent more than female in female-dominated occupations.
- Male earn 20.099 $((e^{0.183144} - 1) * 100)$ percent more than female in male-dominated occupations.
- Male earn 29.106 $((e^{0.2554662} - 1) * 100)$ percent more than female in equal-dominated occupations.

To examine the robustness of the observed wage gap between genders in each category of occupation, we conducted three t-tests. The purpose of these tests was to evaluate the statistical significance of the wage gap estimates and to ensure the reliability of our findings:

Table 6: ANOVA table for testing gender and occupational dominance effects on wages.

Test	F(1, 165710)	Prob > F
MALE + OCC_FEMALE + MALE_IN_OCC_FEMALE = OCC_FEMALE	163.55	0.0000
MALE + OCC_MALE + MALE_IN_OCC_MALE = OCC_MALE	393.38	0.0000
MALE = 0	1905.38	0.0000

By performing three t-tests, we confirmed that the differences in wages between genders in all categories of occupation are significant with p-values = 0.000. Therefore, based on our analysis, it can be inferred that the observed wage gap between genders in each occupation possesses a certain degree of credibility.

Given the limited set of covariates included in our model, it is important to acknowledge that some potential threats to validity may persist, such as omitted variable bias. For instance, marital status is a pertinent socio-demographic variable that may influence female's occupational choices and their wages. As such, the impact of marital status should be taken into consideration when estimating the wage gap between genders. Moreover, individuals' parental education level could also act as an omitted variable in this context. Empirical studies have revealed that higher parental education levels are positively associated with greater educational attainment and higher wages for individuals. Therefore, the impact of parental education on wage differentials could also lead to potential biases in estimating the wage gap between genders. The presence of omitted variables can lead to an incorrect estimate of the relationship between the variables that are included in the model. It can lead to both overestimation and underestimation of the true effect of the included variables on the outcome of interest, which undermines our causal relationship.

5 Conclusion

In this project, we set out to investigate the gender wage gap across various occupations. To do so, we selected variables that seemed to be highly correlated with wages, cleaned the data, defined interaction terms, and conducted regression analysis.

Upon examining the results and testing for significance and robustness, we discovered that the gender wage gap varies significantly across different occupations. Specifically, we found that being a male worker might cause him earn 14.182 percent more than a female worker in female-dominated occupations, 20.099 percent more in male-dominated occupations, and 29.106 percent more in equal-dominated occupations. Interestingly, we observed that the gender wage gap was smallest in female-dominated fields and largest in equal-dominated fields, where it was double the size of the wage gap in female-dominated fields. Our finding could potentially indicate discrimination in wages based on gender. The fact that male workers earn significantly more than female workers in equal-dominated fields suggests that gender biases and stereotypes may still be present, resulting in unequal pay for equal work.

Our project has several limitations that we need to consider. Firstly, there is the issue of omitted variable bias that we discussed earlier, which may lead to underestimating or overestimating the true coefficients. Secondly, all the covariates in our model are categorical variables, and we could potentially improve the precision of our analysis by changing and including some numerical variables. For instance, we could redesign the education level variable to be numerical instead of dividing it into three groups. This would help us avoid losing information by discretizing continuous values. Thirdly, our model produced different results than the empirical differences in means that we observed in the beginning of our analysis. While both approaches show that the largest wage gap between genders occurs in equal-dominated occupations, the empirical differences of mean show that the smallest wage gap happen in male-dominated occupations(much less) while our model shows the smallest wage gap happen in female-dominated occupations. We have not figured out a reasonable explanation for this discrepancy.

If we had more time, we would like to develop a more refined and robust model to better capture causal relationships and investigate the extent to which discrimination may be driving the observed wage gap (Discrimination would be the unexplained factor).

Appendix

The source code and dataset for this analysis could be found on this GitHub repository.

References

Paula England Asaf Levanon and Paul Allison. Occupational feminization and pay: Assessing causal dynamics using 1950-2000 u.s. census data. *U.S. Government Printing Office*, 2009. URL https://www.jstor.org/stable/40645826.

Naomi Cahn. News on the gendered wage gap - and covid-19. Forbes, March 2022. URL https://www.forbes.com/sites/naomicahn/2022/03/31/news-on-the-gendered-wage-gap--and-covid-19/?sh=10f8156c3e4e.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Katherine Meyer, J. David Pacas, and Matthew Sobek. Ipums-cps: Version 7.0 [dataset]. *Minneapolis: IPUMS*, 2020. URL https://cps.ipums.org/cps/.

United States Census Bureau. Income and poverty in the united states: 2020. *U.S. Government Printing Office*, 2021. URL https://www.census.gov/library/publications/2021/demo/p60-276.html.

Megan Wisniewski. In puerto rico, no gap in median earnings between men and women. United States Census Bureau, March 2022. URL https://www.census.gov/library/stories/2022/03/puerto-rico-no-gap-in-median-earnings-between-men-and-women.html.