Analyzing whether the gender pay gap exists using propensity score matching

Yiyun Sun

DEC.20, 2020

Abstract

The gender pay gap has always been a potential problem in society, and hundreds of articles are on this topic. This report also analyzes the gender pay gap, based on data from the Glassdoor website. The propensity score matching method and the linear regression method were used in the complete analysis to obtain the changes in wages under one-to-one correspondent genders. The result indicated that the gender pay gap truly exists, and the working department is another critical factor influencing salary change. All results are worth pondering.

Code and data supporting this analysis is available at: https://github.com/jlkuee/Final-project

Key words

Gender pay gap, propensity score matching, linear regression model, Glassdoor

1. Introduction

The social issue of gender inequality has always been a hot topic. The gender pay gap is one of the core problems among them, which is defined as the difference between female and male employees (The Gender Pay Gap | Wage Gap in Canada | The Facts, 2020). The news written by Sara Ashley O'Brien in 2015 indicated that the median salary of a male teacher is 1,096 dollars a week, whereas the median wage of women is 956 dollars (O'Brien, 2015). Thus, to check if there is a causal inference between gender and salary and analyze how gender impacts salary within each job, governments must adjust the wage properly and alleviate social conflicts.

The most suitable way to make causal inferences in the observational data is matching (Holland, 1986, p. 959). Propensity score matching is a standard matching method and has become welcome in recent years (King & Nielsen, 2018). Thereby, the propensity score matching method is used in this report to ensure that every female is matched with a male in the same condition to eliminate confounding interference. Moreover, a linear regression model is built to demonstrate how salary changed under the influence of different genders.

This report's primary dataset to construct propensity score matching and build regression model is "Glassdoor- Analyze Gender Pay Gap" data, obtained from the Glassdoor website. The detailed description of data will be explained in the methodology – 2.1 data section. Explanation of model and propensity score matching will be in the methodology – 2.2~2.4 sections. Next, all analysis results, including several

tables, graphs about the data, matching result and model result, will be demonstrated in result 3.1, 3.2, 3.3 section, respectively. Relevant conclusions and corresponding weaknesses will be shown in the conclusion section finally.

2. Methodology

2.1 Data

This report's analysis is based on a 2019 online survey on full-time workers provided by the Glassdoor Chief Economist (Chamberlain, 2019), converted to a data – the "Gender Pay Gap" data. The survey aims to discuss the gender pay gap, so the data collected the respondents' essential information, such as gender, age, job title, company department, highest education, seniority level, base pay, and bonus pay. The target population here is all qualified full-time employees, while the frame population is the set of all people who saw this investigation on the Glassdoor website. The sample then consists of the 1000 participants with usable responses randomly selected (Chamberlain, 2019).

Six variables were picked from the "Gender Pay Gap" data: "gender, age, company department, education, seniority" to the analysis, of which only "age" is a numerical variable, and the others are categorical. To make the analysis more straightforward, a new numerical variable, "salary," which means the sum of base pay and bonus pay, was constructed. Considering the purpose is to find the gender pay gap, another new variable, "gender_F," was created. "gender_F" is a dummy variable, which equals 1 when the gender of the respondents is female, otherwise equals to 0. Table 1 provides the respondents' baseline characteristics, a basic summary of the data for analysis. The biggest strength of this data is that, as presented in table 1, it contains approximately one to one female to male ratio, accounting for 46.8% and 53.2%, respectively. The last row of Table 1 demonstrates that females' median annual salary is USD 96571 against a median of USD 105100, indicating the gender pay gap indeed exists, and some other factors influenced. However, the drawback is that the number of data variables is not enough, making the analysis somewhat limited.

Table 1: Baseline Characteristic Summary

Characteristic	**Female**, N = 468	**Male**, N = 532
___age___	42 (30, 54)	40 (28, 55)
___performance_evaluation___		
1	106 (23%)	92 (17%)
2	90 (19%)	102 (19%)
3	88 (19%)	106 (20%)
4	96 (21%)	111 (21%)
5	88 (19%)	121 (23%)
___education___		
College	123 (26%)	118 (22%)
High School	132 (28%)	133 (25%)
Masters	107 (23%)	149 (28%)
PhD	106 (23%)	132 (25%)
___department___		
Administration	95 (20%)	98 (18%)
Engineering	89 (19%)	103 (19%)
Management	87 (19%)	111 (21%)
Operations	96 (21%)	114 (21%)
Sales	101 (22%)	106 (20%)
___seniority___		
1	83 (18%)	112 (21%)
2	102 (22%)	107 (20%)
3	106 (23%)	113 (21%)
4	80 (17%)	104 (20%)
5	97 (21%)	96 (18%)
___salary___	96,571 (80,866, 112,660)	105,100 (87,792, 121,617)

Table 1 shows that the number of males and females differs at each level of variables such as performance evaluation, education level, and seniority. At the same time, these variables, more or less, affect the changes in salary, and they have different effects on salary changes under the influence of gender. Therefore, to reduce the errors caused by these confounding variables, the subsequent analysis will establish a model to eliminate interference factors between groups.

2.2 Multiple linear regression model

Although the purpose of this analysis is to judge whether gender pay gap exists, there are other factors influenced salary changed. In order to explore the relationship between wage and other variables, linear regression model was built using lm() function in R markdown (John et al., 2020). The final model is

$$Y_{salary} = \beta_0+\beta_1 X_{gender}+\beta_2 X_{age}+\beta_3 X_{deptEngineering}+\beta_4 X_{deptManagement}+\beta_5 X_{deptOperations}+\beta_6 X_{deptSales}+$$
$$\beta_7 X_{eduHighSchool} + \beta_8 X_{eduMasters} + \beta_9 X_{eduPhD} + \beta_{10} X_{se2} + \beta_{11} X_{se3} + \beta_{12} X_{se4} + \beta_{13} X_{se5}$$

The salary variable is the dependent variable of the model, which is the respondent's total salary. Among all predictor variables X, the age variable is the only numerical independent variable of this model, which is the respondent's age and others are dummy variables. Gender variable represents the gender of the respondent, equals 1 when the respondent is female and equals 0 when male. Four department variables represent each department; for example, the deptEngineering variable equals 1 when the respondent belongs to the engineering department; otherwise equals 0. Similar for deptManagement, deptOperations and deptSales variables.

The next three variables represent the highest level of education the respondent received; for instance, eduHigh School equals 1 when the respondent's highest education level is high school; otherwise, equal to 0. The last four dummy variables denote the respondent's seniority level, which means if the respondent has worked two years, the se2 variable equals 1; otherwise, it equals 0. Se3, se4, se5 represent the respondent has worked three years, four years, and five years respectively. It is worth mentioning that when the department is administration, all department variables equal 0; when the highest education level is college, all education variables equal 0, and when the seniority level is 1, all seniority variables equal 0.

β_0 represents the intercept of the model and has no real meaning. β_1 represents the mean change of salary while the respondents change from male to female if all other variables remain constant. β_2 is the average difference in salary given a one-unit change in an age when other variables are fixed. The coefficients from β_3 to β_{13} all represent the average difference in salary between the category for which dummy variable equals 0 and the category for which dummy variable equals 1. For example, β_3 denotes the salary gap between the administration department and the engineering department when other variables are fixed.

These independent variables were selected because they are the only variables data provided. Among them, because seniority represents the number of years the respondent has entered the company, it is treated as a categorical variable even though it is a number. The age variable ranges from 28 to 55 within 1000 observations; thus, instead of categorical variables, age is more appropriate as a numerical variable. At the same time, other variables remain their original characteristics.

Besides the model constructed before, two more optional models were made. The first one is

$$Y_{salary} = \beta_0 + \beta_1 X_{gender}$$

, and the second is

$$Y_{salary} = \beta_0 + \beta_1 X_{gender} + \beta_2 X_{age}$$

, named model A and model B respectively. Every aspect of these two models has the same meaning as the final model described previously.

2.3 Propensity score matching

As indicated in the data section, there are other response variables representing respondents' characteristics, such as age and education, besides gender might affect salary change. As a result, the situation discusses the difference between a 42-year-old female's salary, whose highest education level is college, and a 28-year-old male with a master's degree might happen if using the selected linear regression model on the original data. However, this situation does not make sense to the analysis because the female and corresponding contrast male's condition is different, which will cause a random error. Thus, to avoid this situation happening, propensity score matching is needed, which is a methodology to match each female with a similar male.

The main core of propensity score matching is to add probability based on a logistic regression model where the response variable is gender to each observation's values before grouped. The probability here is the propensity score, helping predict the probability of the observation being treated. Then the female group will be matched with the male who has similar propensity score. Therefore, the first step is to estimate the propensity score. Since the response variable gender is a binary variable – gender equal 1 when the sex of respondent is female; otherwise, equal 0, a logistic regression model using glm() function in R markdown was built (John et al., 2020), which is

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{age} + \beta_2 X_{deptEngineering} + \beta_3 X_{deptManagement} + \beta_4 X_{deptOperations} + \beta_5 X_{deptSales} +$$
$$\beta_6 X_{eduHighSchool} + \beta_7 X_{eduMasters} + \beta_8 X_{eduPhD} + \beta_9 X_{se2} + \beta_{10} X_{se3} + \beta_{11} X_{se4} + \beta_{12} X_{se5}$$

$log(\frac{p}{1-p})$ is a logit function, where p represents the probability of the respondent is female, in other words, p is the propensity score.

All predictor variables X have the same meaning as the multiple linear regression model constructed before, and they were explained in detail in the 2.2 Model section. Nevertheless, for the coefficient β, there are some slight differences from before. Coefficients here represent the change in log odds; for instance, β_2 is the difference of the log odds of being female when the age increase by one unit and others are fixed. Furthermore, β_3 represents the difference of log odds of the probability of being female between the administration department and the engineering department when other variables are fixed. Whereas, β_6 represents the difference of log odds of the probability of being female between the highest-level education is college and high school if other variables are constant.

After each observation has its propensity score, it is time to do the matching. This analysis's matching method is the nearest neighbour matching, finding the closest matched male to each female (Alexander, 2020). The final step is to use the selected linear regression model based on the matched data to avoid interference.

2.4 Model check

As indicated in table 2, the final model was determined to be the best based on having the smallest model fit statistics AIC and BIC, with 22088.43 and 22162.05. Moreover, the adjusted R^2 of the final model is the biggest, which is 0.6434, representing the model explains 64.34% the variability of the response data around its mean. Thus, the final model was chosen to study the relationship between annual salary and other factors and will apply to the matched data.

Table 2: Model diagonose

	model A	model B	final model
AIC	23079.74945	22726.8268	22088.4299
BIC	23094.47271	22746.4578	22162.0463
Adjusted R-squared	0.02749	0.3174	0.6434

3. Results

3.1 Data

According to a 2017 gender pay gap report, industries and occupations are the critical factors affecting men's and women's wages (Meara et al., 2017). Table 3 represents how female and male salary changed between different departments in "Gender Pay Gap" data. Salaries in different departments are different, where the management department offers the highest annual salary, USD 99172.13, for females and males received the highest annual salary, USD 108408.48, from the engineering department. However, females' wages are generally lower than those of males in the same department, where the gender pay gap in the engineering department is the highest, nearly 11100.37 dollars, proving the gender pay gap.

Table 3: Gender salary gap between different departments

gender	department	N	average salary
Female	Sales	101	99222.35
Male	Sales	106	108388.42
Female	Operations	96	93098.47
Male	Operations	114	99214.73
Female	Management	87	99172.13
Male	Management	111	106201.32
Female	Engineering	89	97308.11
Male	Engineering	103	108408.48
Female	Administration	95	93429.16
Male	Administration	98	102680.28

3.2 Propensity score matching result

Table 4 demonstrates the summary of the logistic model, which is

$$log(\frac{p}{1-p}) = -0.1940 + 0.0040X_{age} - 0.1117X_{deptEngineering} - 0.2291X_{deptManagement} - 0.1429X_{deptOperations}$$
$$-0.0326X_{deptSales} - 0.0422X_{eduHighSchool} - 0.3789X_{eduMasters} - 0.2619X_{eduPhD}$$
$$+0.2443X_{se2} + 0.2414X_{se3} + 0.0372X_{se4} + 0.3191X_{se5}$$

, as indicated in the table 4.

Among all independent variables, X, all department variables and education variables negatively impact the probability's log-odds, while others have a positive influence. For instance, a unit increase in X_{age} corresponds to a 0.4% increase in the log odds of being female, which means the probability will increase by 50.10% — similar interpretation for other coefficients.

Table 4: Logistic model summary

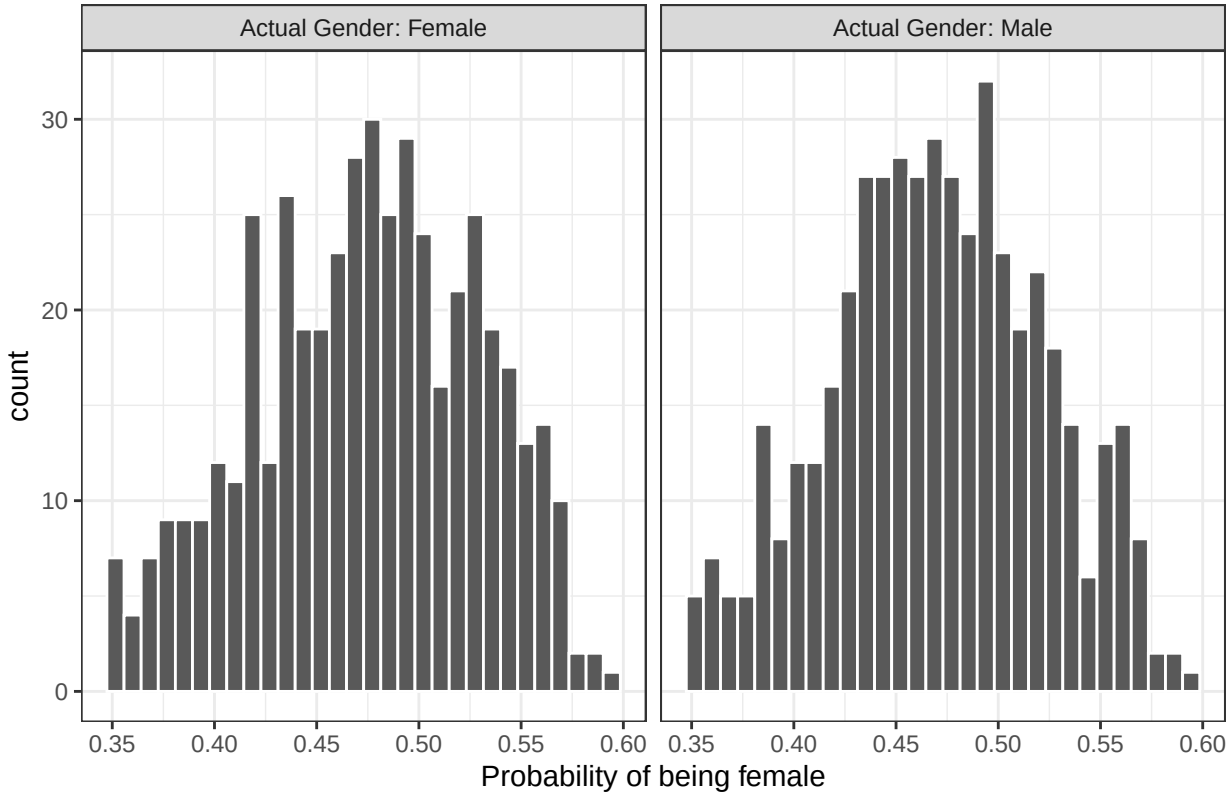
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1940	0.2898	-0.6697	0.5031
age	0.0040	0.0045	0.8954	0.3706
deptEngineering	-0.1117	0.2056	-0.5429	0.5872
deptManagement	-0.2291	0.2049	-1.1181	0.2635
deptOperations	-0.1429	0.2022	-0.7065	0.4799
deptSales	-0.0326	0.2021	-0.1615	0.8717
educationHigh School	-0.0422	0.1791	-0.2355	0.8138
educationMasters	-0.3789	0.1817	-2.0857	0.0370
educationPhD	-0.2619	0.1844	-1.4202	0.1555
as.factor(seniority)2	0.2443	0.2012	1.2139	0.2248
as.factor(seniority)3	0.2414	0.1994	1.2108	0.2260
as.factor(seniority)4	0.0372	0.2091	0.1782	0.8586
as.factor(seniority)5	0.3191	0.2060	1.5489	0.1214

The next step is to match the observations with the propensity score. Table 5 presents the matching result: 936 matched observations and 13 unmatched observations, which means there will be 468 pairs in the matched data. Figure 1 contains two histograms showing the distribution of propensity scores within two gender groups. It is obvious to find two distribution are nearly centred, and the total count for each bar is similar; in other words, the error of establishing a model on this matched data will be small because the female group and the male group are similar enough to ensure one to one correspondence. Table 5 and figure 1 present that the matched data is excellent since the number of observations is as significant as necessary.

Table 5: Matching result

Number of matched	Number of unmatched
936	13

Figure 1: Estimated propensity scores by gender



3.3 Model result

After applying the selected linear model to the matched data, the final model is

$$Y_{salary} = 31522 - 9239X_{gender} + 970X_{age} + 4217X_{deptEngineering} + 5746XdeptManagement + 723X_{deptOperations} + 5974X_{deptSales} - 2093X_{eduHighSchool} + 3301X_{eduMasters} + 6803X_{eduPhD} + 9002X_{se2} + 18181X_{se3} + 29864X_{se4} + 39169X_{se5}$$

, as presented in the table 6. The beta column in the table 6 represents the coefficient of each variable, while 95% CI represents 95% confidence interval, which means there is a 95% chance the true coefficient is between the interval. All p-values are small enough so that it is statistically significant to accept the model. Thus, the model demonstrates that at 95% confidence, women's annual salary is USD 9239 lower than men with the same conditions. Other variables, except the $X_{eduHighSchool}$, all positively affect salary change. For example, when a person grows by one year, the corresponding salary can increase by USD 970 a year. The coefficient of department variables, such as 4217, represents when a person changes from administration department to engineering department and other conditions remain the same, his salary will increase by USD 4217 annually. When the person's highest education level is master instead of college, the corresponding annual salary can increase by USD 3301; but if the highest education level becomes education, the corresponding annual salary will decrease by USD 2093. Similarly, while employees' seniority level changed from 1 to 2, the salary will rise by USD 9002 a year.

Table 6: Linear model summary

Characteristic	**Beta**	**95% CI**	**p-value**
(Intercept)	40,761	36,291, 45,232	<0.001
gender_F	-9,239	-11,159, -7,320	<0.001
age	970	902, 1,038	<0.001
dept			
Administration			
Engineering	4,217	1,130, 7,303	0.007
Management	5,746	2,668, 8,825	<0.001
Operations	723	-2,312, 3,758	0.6
Sales	5,974	2,962, 8,986	<0.001
education			
College			
High School	-2,093	-4,724, 538	0.12
Masters	3,301	559, 6,044	0.018
PhD	6,803	4,034, 9,571	<0.001
seniority			
1			
2	9,002	5,952, 12,051	<0.001
3	18,181	15,156, 21,206	<0.001
4	29,864	26,642, 33,086	<0.001
5	39,169	36,067, 42,271	<0.001

4. Discussion

4.1 Summary

The complete analysis can be divided into four steps. The first step in the 2.1 section is to obtain the "Gender Pay Gap" data from the Glassdoor website and then clean it to make the data meet the following analysis's needs. The data contains six main variables: "gender, age, company department, education, seniority", where "age" is a numerical variable, and the others are categorical. Next is to build a linear regression model to analyze the relationship between the salary and respondent's characteristics, especially for gender in the 2.2 section. To ensure the one-to-one correspondence among different gender, the propensity score matching method was used in the 2.3 section. A logistic regression model was built to calculate the propensity score, and then the data were matched based on the propensity score using the nearest neighbour matching. The last step is to apply the most appropriate linear regression model on the matched data in the 2.4 section.

4.2 Conclusion

In conclusion, there are several findings based on the results. The data summary in the 3.1 section indicated that the salary is different between different gender among each department. Moreover, females' wages are lower than those of males in every department, especially in the engineering department; males' annual salary is 11100.37 dollars higher than females'. After applying the final model chosen on the matched data, the result represents that the female's annual salary is 9239 dollars lower than the annual salary of men of the same age, working in the same department, receiving the same education the same seniority level. Thus, the gender pay gap indeed exists based on the glassdoor data.

Another finding is that besides the gender, if the respondent only gets a high school diploma, whose

annual salary is 2093 dollars lower than the employee with a college diploma. Overall, the government must pay attention to the gender pay gap problem and appropriately reduce it. Considering the education variable, it is also important to provide more learning opportunities for people who cannot go to college. Surprisingly, the average annual salary of females in every occupation is lower than males. However, this data's salary level is higher than the average social wage, so the result may not be that accurate but worth to conclude (Dodge, 2020).

4.3 Weakness & Next step

In terms of the weakness of the data, the variables provided by it are not enough. The author mentioned there are "Race or ethnicity" and "City or state location" questions in the survey, but these two variables disappear in the final data, which might lead to a different result if they exist (Chamberlain, 2019). Because of the loss of the workplace variable, this data cannot indicate which country it is for, which is difficult to conclude that the gender pay gap is just a social phenomenon in a particular country or a global one. Moreover, there are only 1000 observations in the data, which might limit further analysis. Since the data is obtained from the Glassdoor website, the average salary in this data is 100939.8 dollars, which is higher than the social wage level – 52,600 dollars per year (Dodge, 2020). This situation might cause the conclusion to not very realistic. Another weakness is that the logistic model built for predicting propensity score is not as significant as the linear model, so if there is a better model in the future, it can be used in follow-up research.

There are several things to do next. Firstly, some new variables can be added to the data, such as Covid-19 relevant factors. Covid-19 undoubtedly influenced people's life a lot; thus, an analysis of how the salary changed within different gender during covid-19 times is another topic to discuss. Secondly, more observations need to be collected. Data collectors may consider putting the questionnaire on multiple websites instead of only the Glassdoor website. Lastly, it is worth to find another similar dataset about a specific country salary, which could make the conclusion more convincing.

References

1. Glassdoor- Analyze Gender Pay Gap. (2020, September 12). Kaggle. https://www.kaggle.com/nilimajauhari/glassdoor-analyze-gender-pay-gap

2. The Gender Pay Gap | Wage Gap in Canada | The Facts. (2020, August 4). Canadian Women's Foundation. https://canadianwomen.org/the-facts/the-gender-pay-gap/?gclid=EAIaIQobChMIwMHGt5zC7QIVQuHICh1Y4gCHEAAYASAAEgJNdfD_BwE

3. O'Brien, S. A. (2015, April 13). 78 cents on the dollar: The facts about the gender wage gap. CNNMoney. https://money.cnn.com/2015/04/13/news/economy/equal-pay-day-2015/

4. King, G., & Nielsen, R. (2018, November 10). Why Propensity Scores Should Not Be Used for Matching. https://gking.harvard.edu/files/gking/files/psnot.pdf

5. Holland, P. W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association, 81, 945–960

6. Dodge, M. (2020, January 8). The Average Canadian Salary in 2019. Jobillico.Com. https://www.jobillico.com/blog/en/average-canadian-salary/

7. Rich, B. (2020, November 25). Using the table1 Package to Create HTML Tables of Descriptive Statistics. Cran. https://cran.r-project.org/web/packages/table1/vignettes/table1-examples.html

8. Meara, K., Pastore, F., & Webster, A. (2017, March). Is the Gender Pay Gap in the US Just the Result of Gender Segregation at Work? (No. 10673). http://ftp.iza.org/dp10673.pdf

9. Chamberlain, A. (2019, March). How to Analyze Your Gender Pay Gap: An Employer's Guide. https://www.glassdoor.com/research/app/uploads/sites/2/2019/03/GD_Report_AnalyzingGenderPayGap_v2-2.pdf

10. Convert gtsummary object to a flextable object. (2020, October 23). Rdrr.Io. https://rdrr.io/cran/gtsummary/man/as_flex_table.html

11. John H. Maindonald and W. John Braun (2020). DAAG: Data Analysis and Graphics Data and Functions. R package version 1.24. https://CRAN.R-project.org/package=DAAG

12. Alexander, R. (2020, November 5). Telling Stories With Data: Difference in differences. Telling Stories With Data. https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html