Gender Pay Disparity Report

Brennan Hall

June 9, 2021

Overview.

This report describes the work conducted to examine potential unexplained gender pay disparities for a given set of employees. In doing so, I will review the methodologies I chose to utilize and their respective assumptions along with a summary of my findings in conclusion.

Methodologies.

Hypothesis Testing. For preliminary analysis, I considered simple tests of difference of means and medians of the components making up an employee's total compensation as well as other variables that may directly influence pay, such as the probability of promotion per gender. This ruled out any immediate conclusion that gender influences compensation. For this analysis, I considered the mean (or median in the case of total compensation) difference in variables between Male and Female employees and tested whether those differences were statistically non-zero. As total compensation is a linear function of an employee's base rate, regular hours, overtime hours, and bonus, I conducted these tests on all of these variables. Additionally, I conducted a similar test to determine if the rate of promotion was different between genders. In all tests considered, there was no evidence of gender disparity.

Regression Modeling. In order to account for legitimate business factors, I chose to model the variables of interest using various types of regression models. I will not discuss in detail the theory or derivation of my models here. I provide notes and comments within my accompanying code for this purpose. Prior to developing the model, I chose to remove variables I determined were either unnecessary or redundant such as state, zipcode, and Establishment Name (EstNam). Since the establishment name provides the most granular information, and the state and zipcode can be determined from that information, I chose to only retain the establishment name in my models. Similar justifications were made for other variables as well.

Directly modeling total compensation as a function of gender and the other business factors proved unreliable due to the model not satisfying the theoretical assumptions of regression. In response, I chose to rephrase the question from, "Does gender influence compensation metrics" to, "Does compensation metrics influence the probability of being Male or Female". In this regard, I assume the variable of gender within the dataset follows a Bernoulli distribution (it takes a value of "Male" or "Female" with an unknown probability, p) and model the probability of being male as a linear function of the compensation variables and the other business factors. Doing this allows me to model the influence of compensation metrics on gender, meaning how much an increase in base pay will increase the probability of being Male within the given dataset, while taking into account the other business factors' influence on that same probability.

I develop this model and identify the weights to each of the variables that determine the amount of influence each has on the probability of interest on a smaller subset of the data in order to then test the accuracy of the model later. These weights are shown in the accompanying spreadsheet along with a note on whether each can be considered statistically non-zero. Of particular note is that, of the relevant compensation metrics, the weights corresponding to base rate and bonus are the only ones (tentatively) considered non-zero. These weights would be interpreted as a unit increase in one's base rate would correspond to a 50.5% increase in the probability of being classified as Male or a unit increase in one's bonus percentage would correspond to a 27.5% decrease in the probability of being classified as Male.

However, these are not reliable to make a conclusive decision on whether gender pay disparity exists. To make proper a conclusion, I use the model predict the gender of the employees with the remaining subset of data that was not used to formulate the model. With these predictions, I evaluate the accuracy of the model and compare the results to the naive estimate of the probability of being Male within the test dataset. The naive estimate simply being the proportion of Male employees in the dataset. If the modeled probability estimate is statistically not equal to the naive estimate, then we can conclude that the variables we condition on in our model (the compensation metrics and business factors) provide additional information to the probability.

Conclusion.

The result of the proportion test described above lead me to conclude that there is evidence of gender pay disparity within this set of employees. The test indicates the modeled probability estimate of being classified as Male of 52.8% is surpassed by the naive probability estimate of 67.5%. Considering the two weights corresponding to the base rate and the bonus compensation metrics, I would tentatively conclude the disparity slightly favors Male employees; however, further investigation would be warranted.