Closing the Gap: Analyzing Gender Pay Disparities

Jimi Abbott, Alan Liu, and Aditi Raju

April 21, 2023

1 Introduction

Gender pay equity is a pressing topic in modern day society. As students who are going to enter the workforce in a few short years, we would like to know whether the pay disparity is truly reducing over time. Our main focus with this dataset is examining the gender pay gap using several different regression techniques. Furthermore, we aim to examine the effects of other variables, including some that may not be at the forefront of discussion when it comes to thinking about salaries. Ultimately, this project aims to use the aforementioned techniques to quantify and bring to light gender pay gap issues along with exploring effects on salary as a whole. Extracting the trends or factors that contribute to the gender pay gap is necessary to draw attention to the issue and hopefully this awareness will promote positive change.

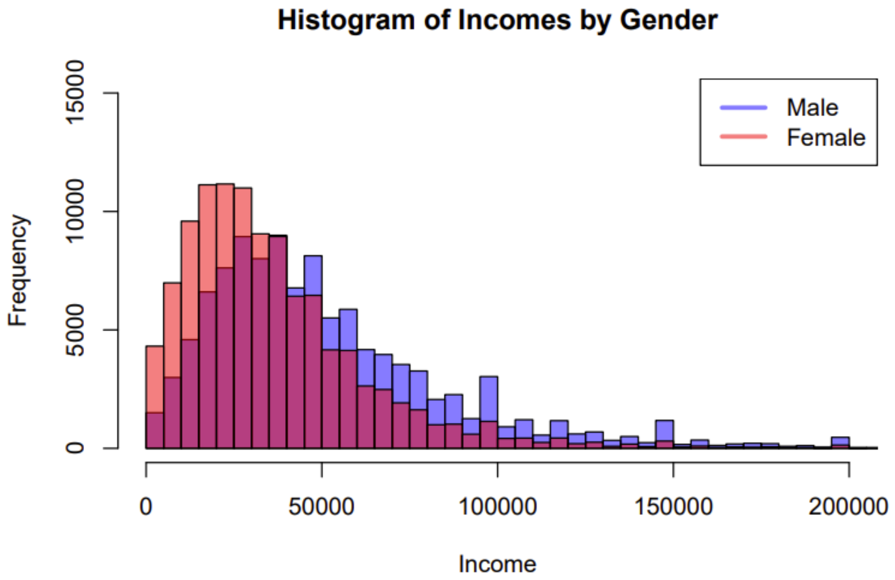
2 Data

2.1 Gender Pay Gap Dataset[1]

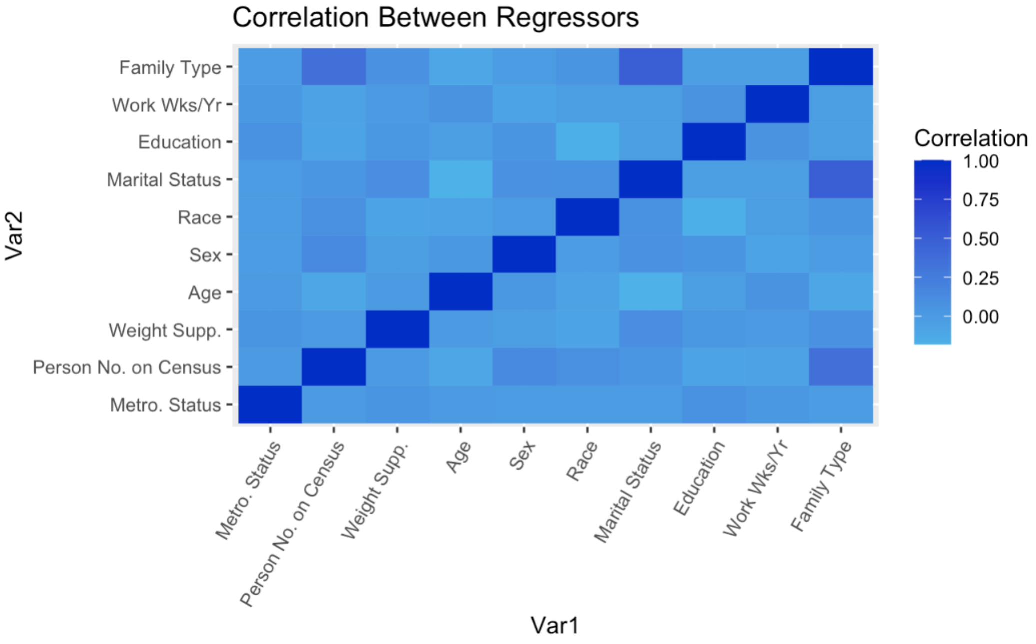
The data set contains census information (Current Population Survey) from the years 1980- 2013, including details about wages, income, industry and most importantly, gender. The data set is relatively large, as it contains 344287 rows and 234 columns. However, most columns contain redundant information like occ, occ1950, and occ1990, which are occupation, occupation on the 1950 basis, and occupation on the 1990 basis, respectively. Furthermore, some of the columns contained NA values, which would affect the regression models so they were removed from our models. The data set is reasonably balanced in terms of our primary factor of investigation, which is gender; approximately 49% of the survey subjects were female. Some regressors of interest included: metropolitan status, age, sex, race, marital status, education level, industry type, full-time or part-time worker, family type, etc. We chose to use annual wage (incwage) as our response variable.

2.2 Visualizations

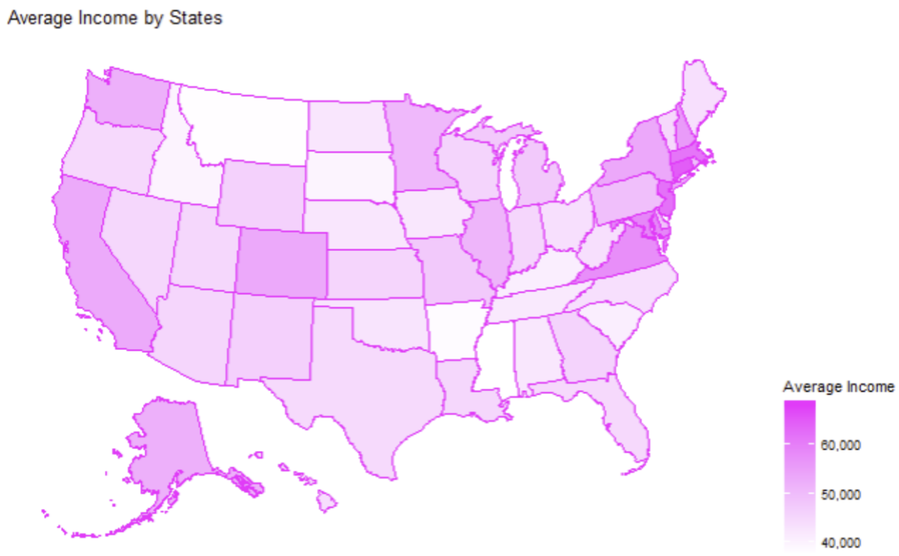
To understand the general trend and distribution of the data set, we used a variety of visualizations.



This histogram is quite representative of the actual distribution of incomes in the U.S. based on its relative log-normal shape. This plot clearly shows the gender pay gap that we aim to investigate: females have a higher representation in lower income brackets, whereas males have a higher representation in higher income brackets - a clear difference between the two genders.



The correlation heat map depicts the pairwise correlation of select regressors of interest in our project. From this, we can confirm that the regressors have low correlation with one another (except for with themselves), which is desired for variable selection to avoid confounding between the model parameters and problems related to collinearity.



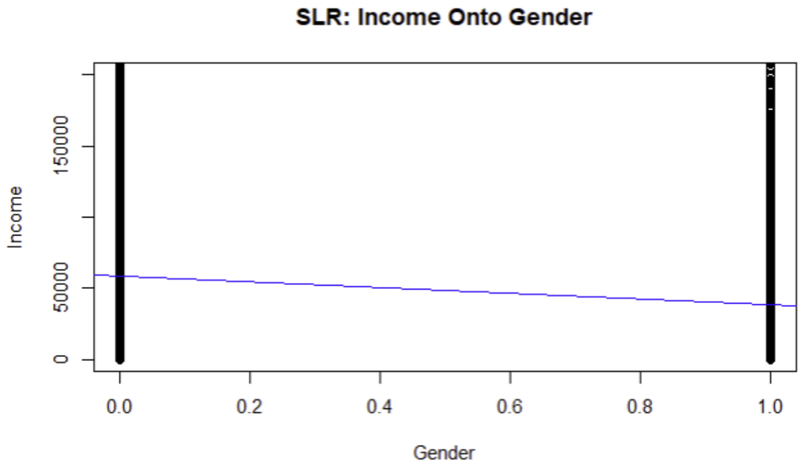
When considering effects on salaries as a whole, it's interesting to note the differences in salaries across different regions of the United states. This map uses the average income by state. On the lowest of lows, we see Mississippi with an average income of $37,611 compared to the District of Columbia, which has an average income of $68,817.

3 Data Analysis and Models

3.1 Simple Linear Regression

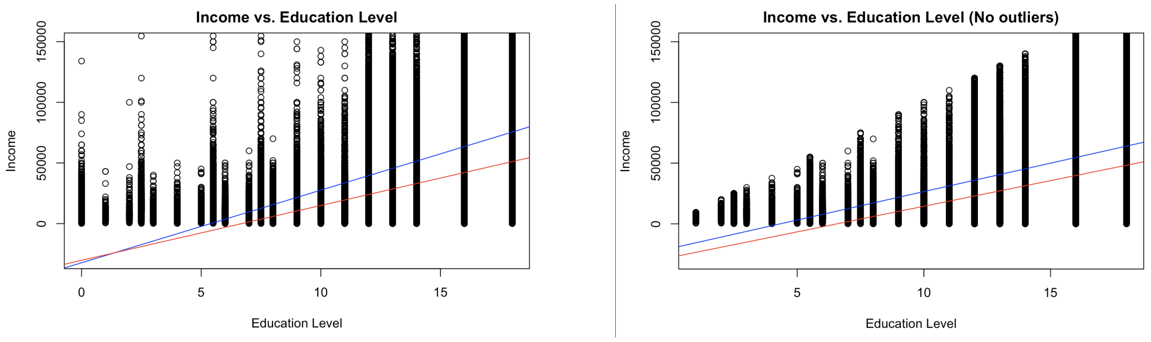
3.1.1 Income on Gender

$$Income_i = \beta_0 + \beta_1 Gender_i + \epsilon_i$$



For our first model, we perform simple linear regression using the recent data (2007 - 2013), regressing salary onto gender with male being denoted as 0 and female being denoted as 1. From our model, we estimate $\hat{\beta}_0 = 58489$ and $\hat{\beta}_1 = -19740$. We test the null hypothesis that the slope β_1, the difference of the incomes between the two genders, is equal to 0, and obtain a t-value of 90.73 and a p-value that's less than $2 * 10^{-16}$.

3.1.2 Income on Education Level



(a) With Outliers

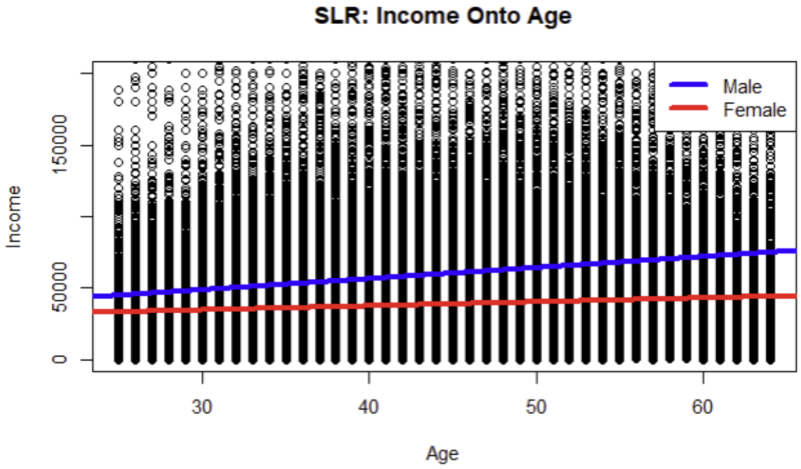
(b) Without Outliers

$$Income_i = \beta_0 + \beta_1 EducationLevel_i + \epsilon_i$$

For our next two models, we perform simple linear regression with education level as the regressor, where education level is a categorical variable representing the grade of education (e.g. sixth grade = 6, eleventh grade = 11, some form of college = 13, etc.). Considering the whole data set, we estimate $\hat{\beta_1}^M = 5990.46$ and $\hat{\beta_1}^F = 4526.11$. However, when removing outliers, a.k.a. observations where income greatly surpasses expectations of the corresponding education level, we obtain new estimates of $\hat{\beta_1}^M = 4686.43$ and $\hat{\beta_1}^F = 4222.27$. This is quite interesting, as we notice the difference between the slopes for males and females with outliers is significantly smaller, which implies males are more likely to exceed their income standards for their educational level, compared to females.

3.1.3 Income on Age

$$Income_i = \beta_0 + \beta_1 Age_i + \epsilon_i$$



This visualization allows us to see the importance of age in relation to income. The slope estimates for male and female, respectively, are $\hat{\beta_1}^M = 779$ and $\hat{\beta_1}^F = 287$. The difference in income per year between the two genders is very noteworthy, and we will use this information later to improve our multiple linear regression model.

3.2 Multiple Linear Regression

We use the income data from 2007 - 2013 to create an MLR model with annual wage as the response. This data contains over 200k observations. A log transformation was applied to annual wage because it was log-normally distributed and residual plots with original model indicated violation of the assumption of homogeneity of variance.

Our model, using both-ways step-wise selection for AIC and BIC, is:

$$log(Income_i) = \beta_0 + \beta_1 log(State_i) + \beta_2 Age_i + \beta_3 Gender_i + \beta_4 Race_i + \beta_5 Race_i^2 + \beta_6 Educ_i + \beta_7 log(Occupation_i) + \beta_8 Industry_i + \beta_9 Wksworked_i + \beta_{10} Hrsperwk_i + \beta_{11} Hrsperyr_i + \beta_{12} Emp_i + \beta_{13} Inflate_i + \beta_{14} Region_i + \beta_{15} Potexp_i + \beta_{16} Gender_i * Age_i + \beta_{17}(Gender_i * Age_i)^2 + \beta_{18}(Gender_i * Age_i)^3 + \beta_{19}(Gender_i * Age_i)^4 + \beta_{20} Relate_i + \beta_{21} Primfam_i + \beta_{22} Nonfam_i + \beta_{23} Privsec_i + \beta_{24} Fedgov_i + \beta_{25} Stgov_i + \beta_{26} Locgov_i + \beta_{27} Nou_i + \beta_{28} Nocov_i + \beta_{29} Union_i + \beta_{30} Met1_i + \beta_{31} Met2_i + \beta_{32} Marpres_i + \beta_{33} Div_i + \epsilon_i$$

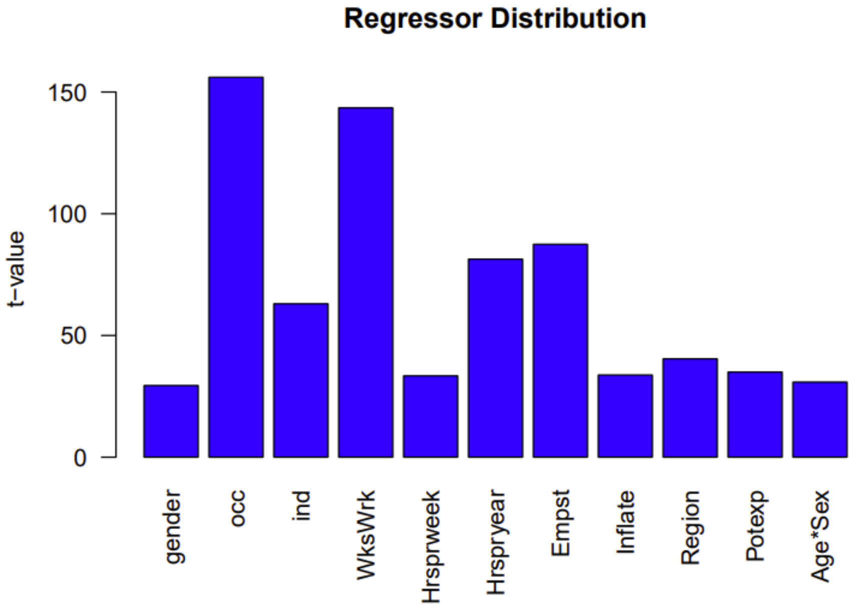
$$\mathbf{R^2 = 0.6592 \text{ and } R^2_{\mathit{adj}} = 0.6591}$$

The income is log-normally distributed, so we take the log of the income and regress it on the rest of the variables. We include the following in the full model:

state, age, gender, race, level of school attained, occupation, industry, weeks worked, hours worked per week, total hours worked, employment status, inflation factor, region, school experience with relation to age, the interaction between gender and age, relationship to head of household, family type, worker type, union status, metropolitan status, and marital status. For family type, worker type, union status, metropolitan status, and marital status, we created dummy variables for each type.

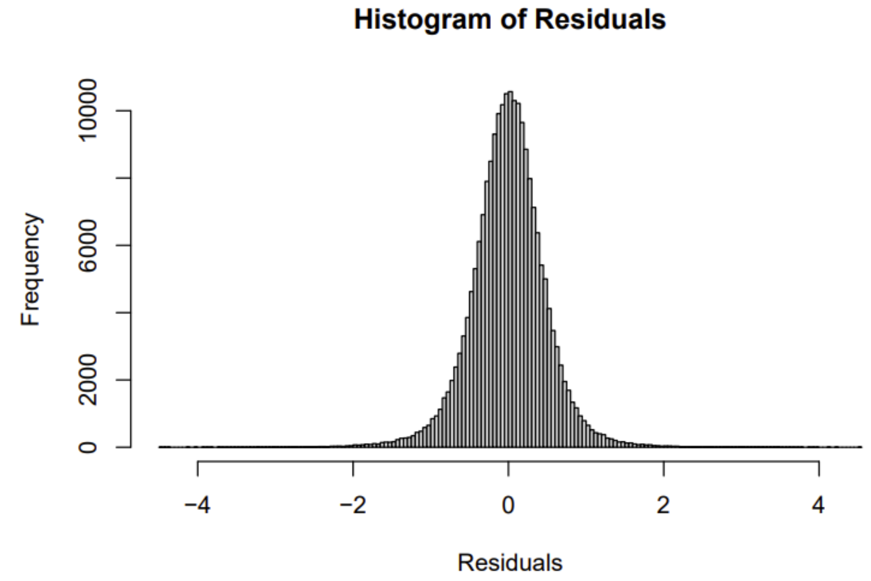
We ran backwards and forwards BIC and AIC stepwise selection, and we are presented with a model that includes each of the previously aforementioned variables. Several of the dummy variables were removed during the process. We are pleased with the effectiveness of our model: looking at the R^2 with this, we say that about 65.91 percent of the proportion of variation can be explained by the regressors.

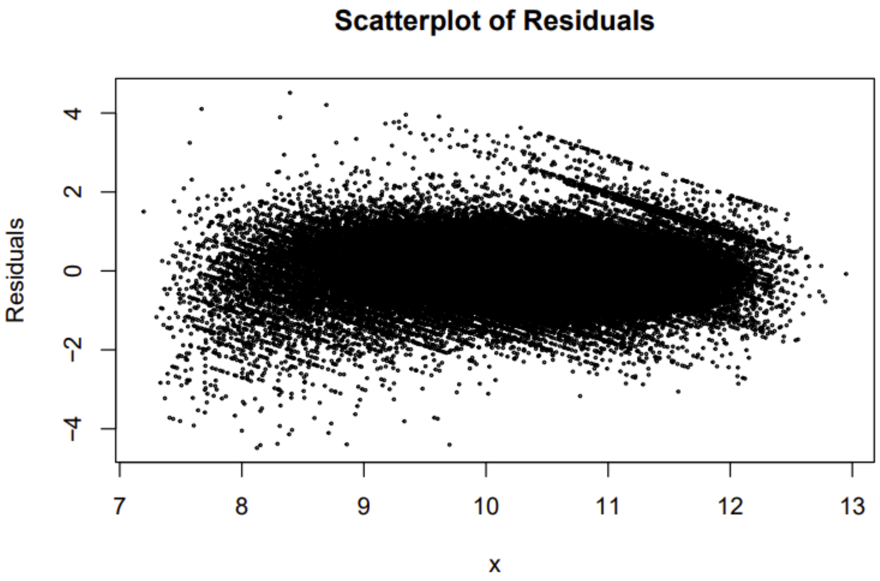
3.2.1 Distribution of Regressors



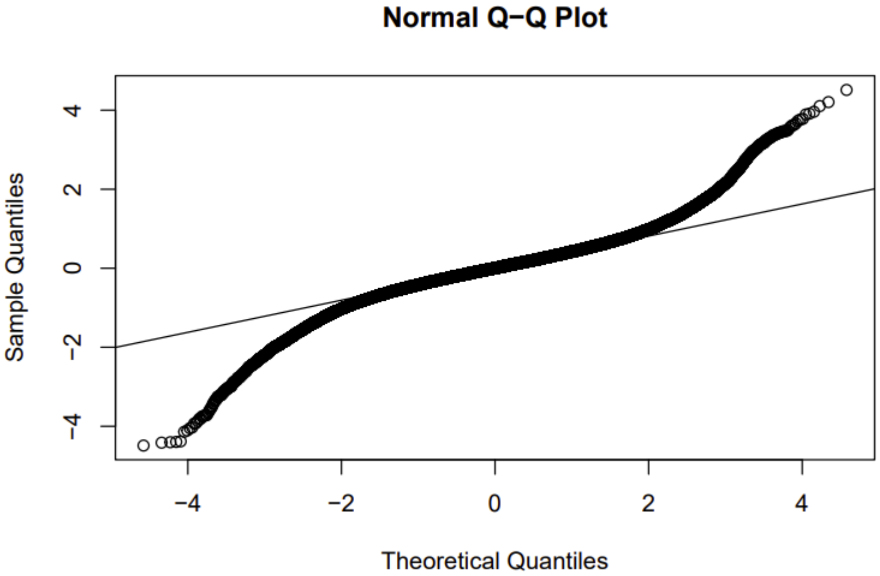
This barplot is a visualization of the 11 most statistically significant regressors compared to the null hypothesis that $\beta_j = 0$. that We see that the two regressors that stand out as being the most statistically significant are occupation and weeks worked per year. In our examination of the gender pay gap, we do notice that gender isn't the most statistically significant out of the lot, but it is certainly important to note that it is still very significant. Overall, we seem to have a lot of statistically meaningful regressors that help us understand the construction of the response, salary.

3.2.2 Residual Analysis





Our histogram of residuals shows that our residuals are normally distributed. However, as we will see in our QQ-plot, that might not necessarily be the case.
The scatterplot with the residuals against the fitted values shows some signs of the violation of the assumption of homogeneity of variance, but if you look at the right side where the plot does look weird, the amount of points compared to the over 200k in the entire plot can be deemed relatively insignificant. Thus, we believe the evidence is not strong enough to say the model violates the assumption of homogeneity of variance.



The QQ-plot shows that our distribution is light-tailed. We believe this is because the proportion of observations that have very high (top 1-2 percent) income or very low (bottom 1-2 percent) income is not equivalent to that proportion for the entirety of the United States. Our data has plenty of observations, but the amount of million dollar or greater incomes is only 74, which is only .0003 percent of all observations. This number is not even close to the real amount, which was estimated recently to be around 0.3 percent.

3.3 Logistic Regression

For our logistic regression, we used many of the same variables from our multiple linear regression model. The goal of this model was to see how accurately we can predict the gender of a person based on all the other data.

$$Gender_i = \beta_0 + \beta_1 log(State_i) + \beta_2 Age_i + \beta_3 Race_i + \beta_4 Educ_i + \beta_5 log(Occupation_i) + \beta_6 Industry_i + \beta_7 Wksworked_i + \beta_8 Hrsperwk_i + \beta_9 Hrsperyr_i + \beta_{10} Emp_i + \beta_{11} Inflate_i + \beta_{12} Region_i + \beta_{13} Potexp_i + \beta_{14} Relate_i + \beta_{15} Primfam_i + \beta_{16} Nonfam_i + \beta_{17} Privsec_i + \beta_{18} Fedgov_i + \beta_{19} Stgov_i + \beta_{20} Locgov_i + \beta_{21} Nou_i + \beta_{22} Nocov_i + \beta_{23} Union_i + \beta_{24} Met1_i + \beta_{25} Met2_i + \beta_{26} Marpres_i + \beta_{27} Div_i + \epsilon_i$$

Trying the model on data yielded an accuracy of 67.6 percent. We believe that this result is significant in the fact that we should not be able to predict gender based off the other variables that easily. In our model, the most statistically significant regressors were total income, occupation, marital status, and relationship to head of household. Income and occupation as two of the most statistically significant regressors in predicting gender certainly shows us some indication of gender pay gap and general unfairness in our society.

4 Summary and Discussion

4.1 Conclusion

Gender does not seem to account for much of the variance in the salary difference overall, but does still play a relatively significant role. For the SLR models, there is a significant difference between the slope of the male and female regression lines. The MLR model is effective in modeling income with an R^2 of 0.659, and while occupation is the most significant regressor in influencing salary, gender is also a statistically significant regressor. It is clear that at the extreme ends of the dataset, where income is low or very high, there is a higher disparity in pay between males and females. Some regressors were surprisingly influential in determining income, like region. The logistic regression model revealed that the factors that were significant in the MLR model were not able to completely predict gender. However, the logistic regression further proved that gender and income have a significant relationship.

4.2 Future Work

In the future, we would like to use more recent data to see whether pay transparency has made a difference in the pay gap. Additionally, the difference in mean working hours between men and women is 5.75 hours a week, which averages to almost 300 hours a year, which could skew the annual income comparison. Instead, gross hourly wage could be chosen as the outcome measure(women are more likely to work part-time) to standardize this difference in number of hours worked per year.

5 References

1. Soriano, Federico. Gender Pay Gap Dataset. Kaggle, 2018,
   https://www.kaggle.com/datasets/fedesoriano/gender-pay-gap-dataset.