Wages and Sex

Efran Himel, Tiffany Quang, Isabelle Hauge, Madelyn Nelson

The “gender pay gap” is a controversial and highly debated topic in the United States. This rises the question amongst people: does sex really affect the a person’s wage? Our project’s goal is to determine how significacntly sex impacts a person’s predicted wage, in 1985. Although, it is likely that the wage disparity between males and females has decreased since 1985, by studying this information we can gain insight that might help make future projections on how this issue may develop. An analysis of the gender pay gap will allow us to gain a better and informed understanding on how such an issue can be tackled and where solutions need to be applied.

In our project, we assessed the impact of sex on wages while taking into consideration other confounding variables. Our data is collected by the 1985 Current Population Survey and contains 11 variables. The four quantitative variables in the data set are level of education, number of years of experience, age, and wage. The remaining seven qualitative variables included in the data are sex, whether or not they live in the south, if they’re in a union, race, occupation, sector, and marital status. Using a multitude of analytical tests, we wanted to determine whether or not each of these variables has a significant effect on wage. We analyzed the distribution among each variable, looked at adjusted residuals, and estimated AIC and BIC to determine the importance of each predictor variable. These methods allowed us to create a regression model for predicting wage, including only the variables we deemed important.

For the first step in our analysis, we plotted each one of our variables against wage to gain an idea of the various distributions that were present within our data. We also produced various histograms for each of our predictor variables as well. Creating these plots gave us a general initial idea of the relationships between each one of our variables in relation to wage. They also gave us a rough sense of which variables had outliers that may be of concern for us later. Overall, from these initial plots we did not encounter any outstanding plots that may provide any interesting insights about our data other than that some relationship does exist between wage on the variables. All our plots had some outlying variables, we interpreted many of these to be unique individuals whose circumstances impacted their wages much differently compared to their contemporaries. For extreme cases, we noted that the outlier may have to be removed when we create our model because they most likely do not represent the general population, such as someone being born into extreme wealth and easily acquiring a high wage job early in life.. For example, in figure 1, our age against wage and experience against wage plots have an extreme outlier in the beginning where they have the highest wage with one of the lowest age and education.

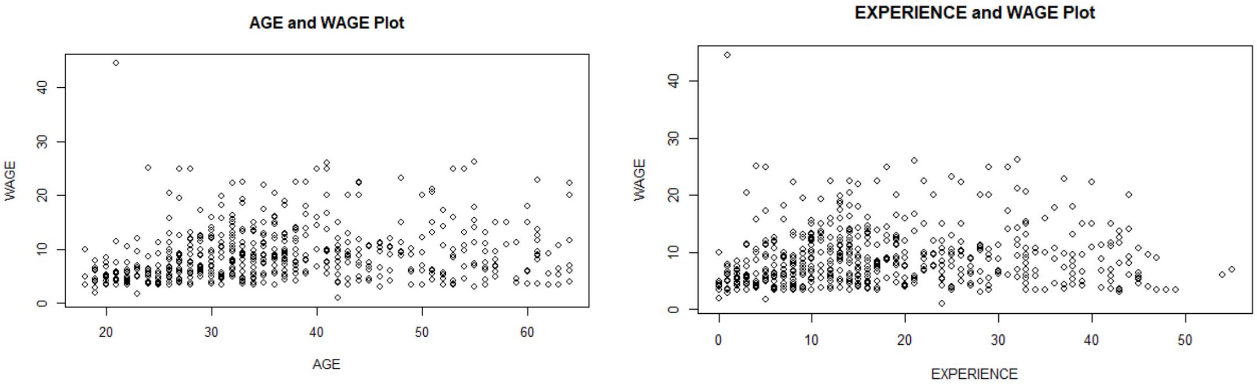

Figure 1: Age and Wage and Experience and Wage Scatterplots

After creating these plots, we produced regression lines for each one to determine which plots had some linear relationship with wages. We would later use the residuals from these lines for my in depth analysis of our data later on.

After our initial data exploration, we continued our analysis with residual plots to see if there were any unusual behaviors. As can be seen from the plot on the left in Figure 2 below, which shows our full data set, there is one outlier that is significantly different from the rest of the data. We removed the outlier and replotted the residuals, which can be seen on the plot to the right.

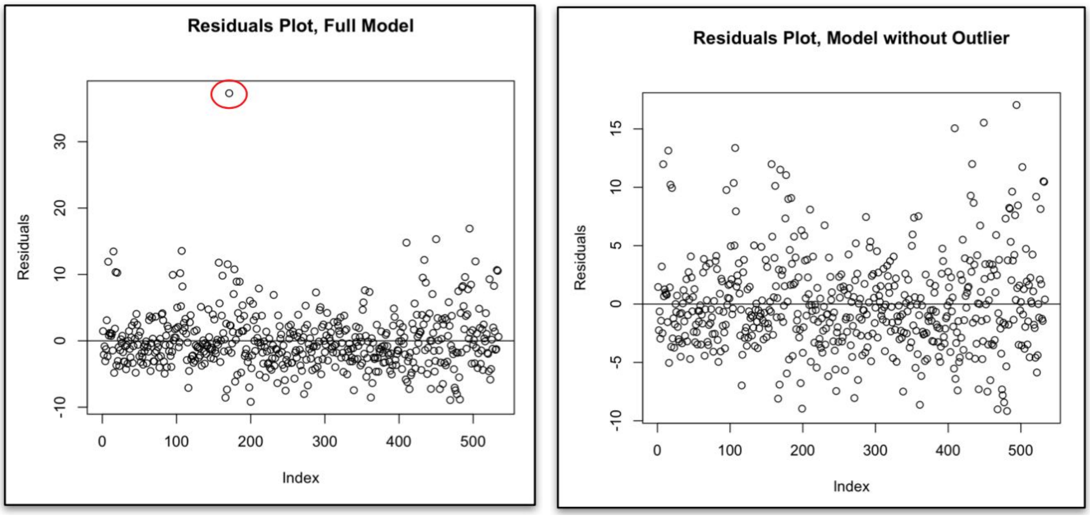

Figure 2: Residual Plots With and Without the Outlier

This outlier represents a white woman age 21 with one year of experience and a wage of 44.50 dollars, which is roughly 18 dollars higher than the next highest wage of 26.29 dollars. We decided that if testing revealed that it would be beneficial to remove this outlier, it would be acceptable to do so because of the fact that that data point is not very representative of the rest of our data set.

During our data analysis, we used the pairs function in R to plot the correlations of each of the variables, and we found that age and experience have a very high correlation of 0.9779, which can be seen in Figure 3 below.

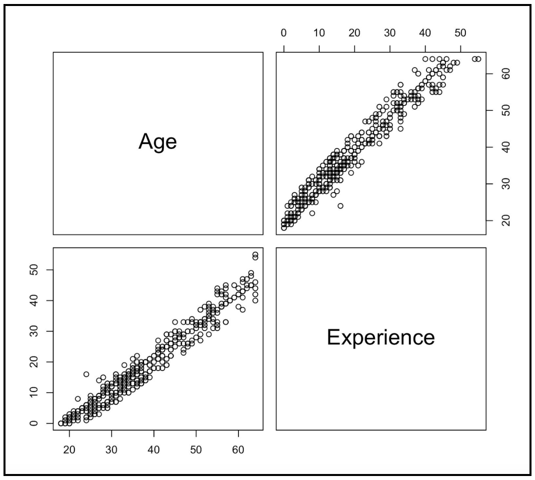

Figure 3: There is a Strong Correlation Between Age and Experience

Due to this fact, we decided that our final model would only include one of these in an effort to eliminate any unnecessary variables, but we had to do more testing to determine which of these variables would remain in our final model.

Q-Q plots of each of our quantitative variables were examined, as well, to make sure that the normality assumption was held. We examined the normality for experience, education, age, and wage in two types of Q-Q plots, one where sex was not considered and another where sex was considered. In general we found that experience, education, and age all had some semblance of normality or almost normality for both types of graphs. However, during our testing, we found that our wage variable showed a right skew, which can be seen from the graph on the left in Figure 4 below.

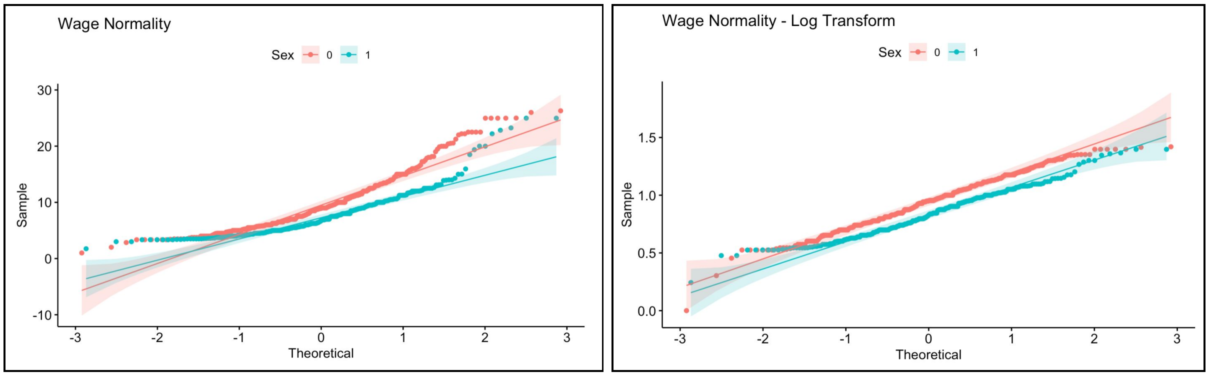

Figure 4: Q-Q Plots of Wage Before and After a Log Transformation (0 = Male, 1 = Female)

We decided to try to fix this using a log transformation, which can be seen from the second Q-Q plot in the figure above. We noticed that this new plot showed much less of a skew, so we decided that a log transformation of the wage variable was something we would want to consider when carrying out further testing.

Another approach we utilized to analyze our data were BIC and AIC tests. This allowed us to test for every possible combination of predictor variables so that we may gain some insight on which variables had the highest predictive capacity for wages in our given data. After completing both tests, we found that all top scoring models in both the AIC and BIC results had sex and union status and some combination of education and experience. We interpreted this finding to mean, that sex and union had the strongest predictive capabilities for our data set in determining seeing what an individuals wage would be.

After gaining these results, we decided to test the robustness of the models. We conducted 10 fold cross validation for all the top ranking models to see how robust these regression models were. All our models had very low R-squared values when run through cross validation, with the best mean R-squared value for our BIC model as 0.286095 and 0.2855159 for our AIC model. After conducting these tests we now believed that sex likely has some relationship with wages, due to it being in all our best models, but does not alone make significant contributions to predicting the final outcome of an individuals wage.

After our data analysis, we had two things that we wanted to consider when building initial models to ensure the linearity assumptions were met and to make sure the models would have a satisfactory predictive ability: the removal of the outlier and the log transformation of the wage variable. In an attempt to determine which of those steps should be taken, we performed adjusted R-squared and BIC testing with the four sets of data we were considering: (1) the original data set, (2) the data set with the log transformation of wage, (3) the data set with the outlier removed, and (4) the data set with the outlier removed and with the log transformation of wage.

We conducted adjusted R-squared testing first, and the four resulting plots can be seen in the figure below:

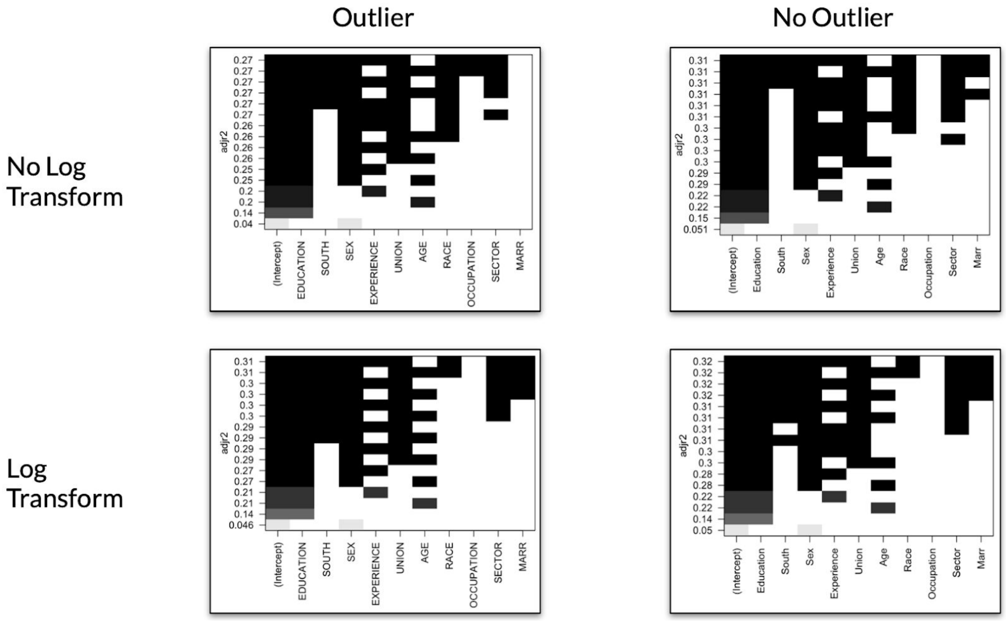

Figure 5: Results of Adjusted R-Squared Testing

With these graphs, it can be seen that the models without the outlier have a higher adjusted R-Squared, which suggests that leaving out the outlier would produce models that were a better fit for the rest of the dataset. As aforementioned, we feel comfortable with removing this point because it is not very representative of the rest of the data. We can also see from Figure 5 that the models with the log transformation of wage outperformed those without the log transformation, suggesting that there is evidence that we should transform the wage variable in our final model.

In carrying out BIC testing, we used the same four sets of data as we did with adjusted R-Squared, which is illustrated in Figure 6. It can be seen that this form of testing returned very similar results to those obtained with adjusted R-Squared testing in regards to the fact that the best fit models had no outlier as well as the log transformation of the wage variable.

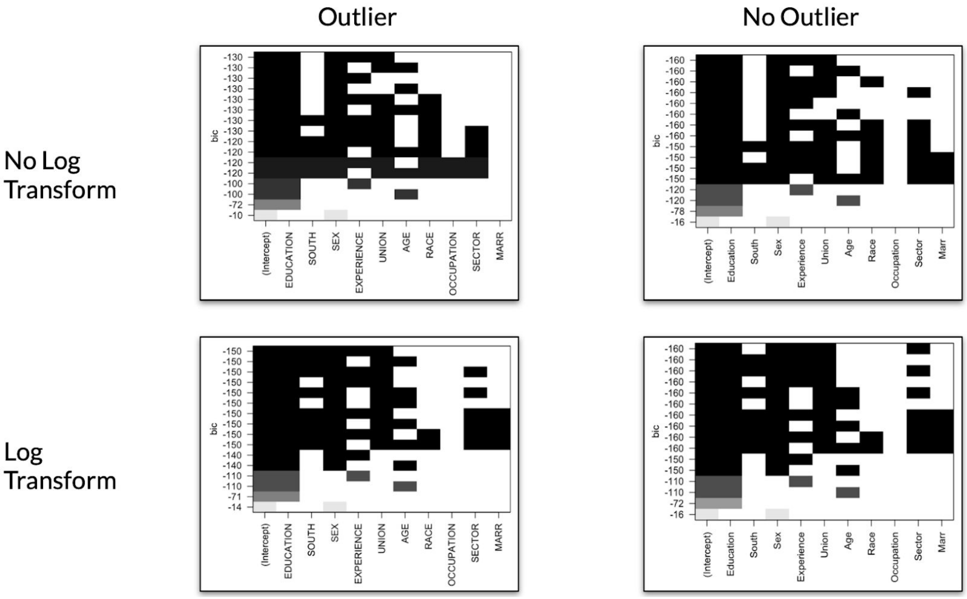

Figure 6: Results of BIC Testing

Therefore, after testing with both adjusted R-Squared and BIC, we decided to remove the outlier from our dataset and do the log transformation on wage in order to uphold the normality assumption.

With the outlier now removed, we wanted to see if our initial model (without outlier removed and no transformations) still upholds. From the new BIC tests and adjusted R-Squared tests, it was determined that the variable sector could be added to our model. Though, we were not sure whether we should include an additional variable to our final model or not as it could make the model become overfit. We performed an ANOVA test at an alpha level of 0.05 between our two models: the initial model and the initial model with sector. Our null hypothesis stated that our initial model without sector is a better performing model. The ANOVA test gave us a p-value of 0.01175, thus we rejected the null hypothesis and stated that sector was a significant variable in our model.

Due to the multicollinearity between age and experience, we had to decide which one of the two variables to omit from our final model. We computed the Variance Inflation Factor for two different models, one that included age and another that included experience. The VIF values for the model that contained age were overall lower and had values closer to one when in comparison to the model that contained experience. Therefore, we decided to include age over experience in our final model.

With the transformation performed, the outlier removed, and an additional variable added to our model, we've decided that we exhausted all possible diagnostic testing and have concluded that we've reached our final model. We ran an analysis on our final model and received

$$\log(\text{wage}) = 0.5009 + 0.01195(\text{age}) + 0.08862(\text{education}) - 0.2211(\text{sex}) + 0.2037(\text{union}) \\ + 0.091908(\text{sector})$$

as our multilinear regression model for determining wage with an adjusted R-squared of 0.3067.

Based on our data from the 1985 Current Population Survey, we found that sex is an important predictor variable in our model to predict wage. Using log transformations of our data to address a lack of normality, we were able to produce a final model for predicting wage that includes the predictor variables age, education, sex, union, and sector. In this model, sex has a negative coefficient meaning it will decrease predicted wage when the value of sex is one, which represents females. Therefore, it is likely that being a female could have a negative impact on wage. Although these findings are interesting, we are unable to make sweeping general statements due to the limitations of our data and tests.