Effects of Job Training on Wages

Team 4

10/7/2021

Summary

Inferential questions regarding our data analysis were based on the two methods used for analysis. A Logistic Regression method focused on **whether or not workers who receive job training tend to be more likely to have positive (non-zero) wages than workers who do not receive job training**. While a Multiple Linear Regression method evaluated **the impact of job training on workers to see who tends to earn higher wages**. The analysis explores quantifying the effect of job training on the odds of having non-zero wages, seeking a likely range for the effects of training, and discovering supportive evidence on how the effects of job training differ by demographic groups. Both Multiple Linear Regression and Logistic Regression were used as methods for analyzing the dataset for information to support answers to our inferential questions. Through our analysis we learned that receiving job training has a statistically significant positive impact on the odds of having positive wages in the future. Workers who received treatment (job training) will also earn an average of $2,573.00 more than workers who did not receive treatment , ceteris paribus.

Introduction

Job training can provide employees with useful knowledge that helps them to improve working skills; however, it also costs money and time. Whether the job training deserves a worker's time, which is represented by the change in the worker's wages, is going to be explored in this paper.

In the 1970s, researchers in the United States ran several randomized experiments intended to evaluate public policy programs. One of the most famous experiments is the National Supported Work (NSW) Demonstration, in which researchers wanted to assess whether or not job training for disadvantaged workers had an effect on their wages. Eligible workers were randomly assigned either to receive job training or not to receive job training. Candidates eligible for the NSW were randomized into the program between March 1975 and July 1977.

A subset of the original dataset from the NSW Demonstration containing only male participants is analyzed in this report. Although the original data is based on a randomized experiment (which means we would have been able to make causal statements directly), the data provided for this project only compares a subset of the data. In the data provided, the treatment group (those who received the training) includes male participants within Lalonde's NSW data for which 1974 earnings can be obtained, and the control group (those who did not receive the training) includes all the unemployed males in 1976 whose income in 1975 was below the poverty level. This control group is based on a matched Current Population Survey - Social Security Administration file.

The rest of the report is organized as follows. Data Part 1&2 describe the data and an exploratory analysis of this data. Model Part 1&2 show the descriptions of our models fitting and selection. Analysis Part 1&2 include the interpretation and our main findings. The Conclusion generalizes our investigation and potential limitations of our research.

Part 1

Data Part 1

We are interested in whether workers who received training tend to earn higher wages than workers who did not receive training. In order to eliminate biases that people has different wages before training, we used the difference between real annual earnings in 1978 and real annual earnings in 1974 as the response variable, denoted as **response**. We checked the distribution for the outcome variable 'response' using a histogram, and the response variable generally follows a normal distribution as the histogram exhibits a bell-shaped curve centered around 0.

For numeric variables age and education, we first centered those variables and created scatter plots with the response variables. For age, there is a positive correlation between response and age. It aligns with our assumption that older people are more likely to earn more income. It is interesting to find out the slightly negative correlation between education and response that people with few years of education are actually more likely to have higher income. We found this finding very counterintuitive.

Then, we discovered interaction effects using boxplots and we did not find obvious differences on response between treat and other demographic variables. We think there might be no interaction effect between treatment and other demographic variables. While looking at the scatter plot to find out the interaction effect between age and response by job training, we find something interesting. For workers who received job training, there is a positive correlation between age and change in income; however, for workers who did not receive job training, there is a negative correlation between age and change in income. Therefore, we believe there is an interaction effect between age and treatment and we might want to include that in our final model.

For categorical variables, we draw box plots to investigate the possible relationships. According to the plot, people who do not have training have a response mean centered around 0 while people with training have a mean above 0. People who were not married had a higher mean than those that were married. The response mean for hispanic workers looks higher than that of not hispanic workers. When checking whether there might exist interaction effects between treat and other factor variables using the anova test, we find nothing significant.

Model Part 1

To begin with, all the numeric variables are mean centered to help avoid potential multicollinearity problems. The baseline model includes all variables available in the dataset as predictors. All four assumptions of linear regression are checked for this model. Points generally follow a random pattern in the residual plots and spread with similar variances, which thus satisfies the linearity, independent and equal variance assumptions. The majority of points lie on the 45-degree line in the QQ-plot except for the tails, so normality assumption is not violated either. Therefore, no transformation is required for this analysis.

The null model includes only treat since it is our variable of interests and the adjusted R Square for the null model is 0.026. The full model consists of all variables individually and all possible combinations of interaction, which is age, education, treat, black, hispan, married, nodegree, and all interactions combinations between them. The adjusted R-Square for the full model is 0.076. We first selected our model using stepwise selection with AIC, the result model consists of treat, age, married, the interaction between treat and the interaction between age and married. We then performed another stepwise model selection with BIC, the result model consists of treat, age and the interaction between treat and age. Since we are interested in the effect of treatment with other demographic variables, we decided to use the result from AIC to keep more variables and interactions. The final model has an adjusted R-Square of 0.072 and has treat, age, married, the interaction between treat and the interaction between age and married. Because the interaction between treatment and demographic groups, which is our question of interest, is dropped during model selection, we ran another F-test to confirm whether it's safe to exclude them. It turned out that we can exclude those interaction terms as the p-value of F-test is greater than 0.05. Therefore, the final model is as follows:

$$Response_i = \beta_0 + \beta_1 treat_i + \beta_2 age\_c_i + \beta_3 married_i + \beta_4 treat : age\_c_i + \beta_5 age\_c : married_i + \epsilon_i$$

$$N(0, \sigma^2), i = 1, \ldots, n$$

The model assessment was then performed on the final model to test whether the final model satisfies all four assumptions for linear regression. Within the 'variables vs residual plots,' points are randomly spread throughout the graph, and it means that the regression coefficients have already captured the relationship, so the linearity assumption is satisfied. For the 'residuals vs fitted' graph, points also spread out randomly, and the trend is flat with equal distances within points, which means the independence and the equal variance assumptions are satisfied. Last, most points lie on the 45-degree line in the Normal QQ plot; therefore, the normality assumption is satisfied.

By calculating the vif, vif for treat is below 2, means that there is no serious multicollinearity issue.By accessing the 'residuals vs leverage' graph, all points lies inside inside the 0.5 cook distance line, indicates that there is no outlier, leverage and influential points in the model.

Part 2

Data Part 2

Since we are interested in whether workers who received job training are more likely to have positive wages than workers who do not receive job training, and the dataset mentioned that the training was taken place between March 1975 and July 1977, we used re78 which is the real annual earnings in 1978 as the response variable. A dummy variable 're78_fact' is created to denote if wages in 1978 is positive. One of the interesting points we found in the dataset is that some people had income in 1974 and 1975 but had 0 income in 1978. In addition, some people had higher income in 1975 but had lower income in 1978. The exploratory data analysis is then performed on the modified dataset to determine trends before proceeding to model selection and significant tests.

(table 1 code) -> INSERT THE CODE CHUNK HERE TO MAKE CREATING FINAL FILE QUICKER

Table 1: training vs. income The table below shows that people without training had a higher probability of having an income. Specifically, people who did not receive training had 2% higher probability of having an income in 1978. This result does not align with our assumption that people with training should have more income, and the reason can be further examined through the model. To determine whether this difference is statistically significant, we need to present a further regression test.

Differences in the probability of having an income are also observed across different race groups. People with race black tend to have a 9% lower probability of having an income in 1978 in comparison to people of other races, while hispanic people have a 8% higher chance of having an income in 1978 in comparison to non-hispanic people. In addition, married people have a slightly higher chance to have an income in 1978 than people who are not married. People with a high school degree also tend to have a higher chance of having income than people without a high school degree although the difference is not very significant.

For numeric predictor variables, the median age does not differ significantly for people who had an income or not, but the range of those who had an income during the year 1978 was larger than the range of those who had an income. Among those who did not have an income, there was a larger proportion of people in the 30-40 age group.

(boxplot for age code) -> INSERT THE CODE CHUNK HERE TO MAKE CREATING FINAL FILE QUICKER

In addition, people who had income in 1978 had a median of education for 11 years, and on the contrary, people who did not have income had a median of education for 10 years.

(boxplot for age * training) -> INSERT THE CODE CHUNK HERE TO MAKE CREATING FINAL FILE QUICKER

We found it interesting that the effect of age on income in 1978 differs by receiving training or not. Among people who did not receive training, people who did not have an income generally had an older age. While for people who received training, the median and the range of age does not have a significant difference between people who had an income or not.

Model Part 2

The baseline model includes age, treat, married, hispan, black, nodegree and education as predictor variables. We then checked the binned residual plot for each of the predictor variables to see if any transformation is required. Within the residual binned plot for age, there is a downward quadratic trend. Thus, we tried to incorporate a quadratic term for age; however, the AUC of the model does not improve. We also tried cube transformation on age but the result did not improve much. For the sake of clearer interpretation, we decided to only incorporate the original age term in the following models. For other binned residual plots, the majority of points lie in the 95% confidence interval. Based on the binned plot between education and income from EDA, there is a gap between below 8 years of education and above; therefore, we decided to transform education into a binary variable with two levels: less than or equal to 8 years and above. After re-fitting the model with the new binary education variable, the AUC improves from 0.618 to 0.635 and the education variable appears to be significant in the regression model. Therefore, we will keep this binary variable for education instead of the numeric variable for future models.

To further investigate if the interaction between training and demographic groups is significant in predicting the odds of having an income, we performed change in deviance tests for each interaction, respectively. The results suggest that only the interaction between training and age is statistically significant given the p-value of Chi-squared test is smaller than 0.05. Additionally, the interaction between training and hispan is only statistically significant at 10% significance level.

The stepwise selection using AIC as criteria is used for model determination. The null model consists of our variables of interest which is treatment. The full model consists of all variables individually and all possible combinations of interaction within the modified dataset. The resulting final model contains variables of treatment, age, black, educ, re74, the interaction effect between treat and age and the interaction between re74 and black.

To ensure it's safe to exclude the interaction terms between treatment and demographic groups, which were originally our questions of interest, from the final model, we performed a deviance test to examine if this interaction effect is significant. According to the ANOVA test, the additional interaction term is not statistically significant. Hence, we can conclude that the effect of training does not differ by race groups. Our final model is as below:

$$y_i|x_i \sim Bernoulli(\pi_i);$$

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0+\beta_1 treat_i+\beta_2 age(centered)_i+\beta_3 black_i+\beta_4 re74_i+\beta_5 educ(binary)_i+\beta_6 treat_i \times age_i+\beta_7 re74_i \times black_i$$

The model assessment was then performed on our final model to test whether the final model satisfies assumptions for logistic regression. Based on binned residual diagnostics, points generally fall within the 95% confidence interval and we do not observe clear patterns in the plot. It means that the relationship has been already captured by the regression coefficients and our final model is reasonable to use. By checking variance inflation factors, we do not observe any multicollinearity problems in our final model as well.

Analysis

Analysis Part 1

According to the results, the p-value of the coefficient for binary treatment variable is smaller than 0.05, indicating that the job training is statistically significant in this model. In other words, workers who receive job training will have 2573 (dollars) more in increase of income on average in comparison to workers who do not receive job training, holding other variables constant. To quantify this effect, the 95% confidence interval for workers who do not receive training is (1143.31, 4002.63), which means the likely range for the effect of job training at 95% confidence interval is between 1143.31 and 4002.63 dollars. It's interesting to note that we have interaction terms for both age versus treatment and age versus married in the final model, which means that the effect of age on the changes of income differs by whether people had training or not, as well as by whether people were married or not. The interaction term between treatment and age is also significant, suggesting that the association between age and changes in wages is different for people who received training and who did not receive training. For people who did not receive training, one unit increase in age is associated with a 72.5 (dollars) decrease in changes of wages. For people who received training, one unit increase in age is associated with a 222.5 (dollars) increase in changes of wages. Furthermore, the associations between age, married or not and the change in wages are also worth noting. Both age and the binary variable which indicates married or not, as well as the interaction term between married and age are found to be statistically significant in the final model. This illustrates that the effect of age on the changes in wages also differs by whether workers are married or not. Last but not least, the adjusted R-square of the final model is 0.072, which indicates only 7.2% of the variation in the changes of wages is explained by this model.

Analysis Part 2

After exp:

The p-value for treatment in the final model is 0.035 which indicates that receiving treatment is a significant factor for positive income in 1978 only at the 0.05 significance level. The odds of people who did not receive training to have income in 1978 is 0.21 times of people who received training, and the 95% confidence interval is between 0.05 and 0.88. The result does not align with our EDA that people who did not receive training had a slightly higher probability to have income in 1978, and it might be because in the regression we controlled all other variables. The binary variable of education is also significant at 0.1 significance level. One of the interesting associations we found in the final model result is around the binary education variable. The odds of people who had less than 9 years of education to have income in 1978 is 1.52 times that of people who had more than 9 years of education, which indicates that people with fewer years of education will have higher odds in having a positive income. The 95% odds confidence interval is between 0.96 and 2.4. The result does not align with our EDA as well as the intuition that people with higher education have higher probability of having income in 1978, and this might be due to the imbalance of the amount of data we have in those two education groups. The interaction effect between training and age is significant, which means the effect of training on having income differs between each age. The interaction effect between re74 and black is also significant. The accuracy of this model is 48.8% and the AUC we obtained is 0.662.

Conclusion

Conclusions Part 1

In conclusion, the most important finding of our study is that workers who received treatment will earn 2,573 (dollars) more on average in comparison to workers who did not receive treatment, holding all other variables constant. The likely range for this effect is dollars (1143.31, 4002.63) with 95% confidence level. Using this result, we should promote more benefits of job training to the public so that workers can earn more income. Age and being not married also have a positive effect on earning more income. In addition, the effect of receiving treatment on change in income also differs within each age group; therefore, we might want to promote different types of treatment to each age group to maximize the effect of job training.

Conclusions Part 2

According to the analysis of job training impact on the probability of getting positive wage in 1978, receiving job training has a statistically significant positive effect on the odds of having positive wages in the future. The odds of people who did not receive training to have income in 1978 is 0.21 times of people who received training. The likely range of the odds ratio is (0.05, 0.88) with 95% confidence level. Additionally, the effect of age on the odds of having a positive income differs by receiving job training or not. It's interesting that the interaction term between wages people had in 1974 and whether people are of race black is also significant, which suggests that the effect of wages in 1974 is different for people of race black and other races.

Limitations

The final models and the analyses reported above do suffer from some drawbacks however:

1. The data for workers above the age of 30 is very sparse, disabling us from understanding how workers in that age bracket respond to treatment.

2. The data that we're using isn't randomized - we would ideally want to compare people who had below poverty wages in some year and then some of them were given training and the others were not. In our case the data may be biased because the treatment group is workers for whom 1974 wages could be obtained and the control group is workers for who had below poverty level wages in 1975.

3. There are 51 workers in the dataset who had a positive income in 1974 and 1978 and no income in 1975. This does not make sense since the difference for some workers between their 1978 and 1974 wages is very extreme. This might be affecting our model in ways we don't understand.

4. There are some data points which are being counted twice due to the categorical nature of the predictors black and hispanic. If a worker is white, the same information is being counted twice because both black and hispanic would be 0 in this case. A better way to encode this information might be to make one predictor race which takes different values for different races so that the same information is not counted twice.

5. The data set is imbalanced. For example, the sample size of the treatment group is only about one-third of the control group, the majority of people had more than 8 years of education after we collapsed the education variable, which could potentially tweak the results.

6. The R-squared and accuracy from the two models above are both not high, which indicates a very limited predictive power. There might be other potential variables that may influence the wages of workers not contained in the dataset, such as workers' performance during the job training.