Team proj 1

Team 6 purple

9/17/2020

Part I

1. Introduction

It is common to believe that job training is efficient in boosting the earnings of disadvantaged workers. In the 1970s, several experiments were conducted to explore the real impact of job training on wages, including the National Supported Work (NSW) Demonstration[1]. In the experiment, eligible workers were assigned to receive job training, and their incomes in 1973, 1974, and 1977 were recorded accordingly.

This primary goal of this analysis is to assess whether receiving job training has a significant effect on annual earnings and what is the range of the impact. Other considerations include whether this effect differs by demographic groups in terms of age, education, and racial identity. We're also interested in exploring other associations with wages.

2. Data

2.1 Data Pre-Processing The original data set is accessed via the NSW Demonstration experiment with 614 non-empty observations. To explore the effect of getting job training on annual wage, we create a numeric variable 'change' to quantify the changes in earnings from 1974 to 1978. By quantifying the response as the change in income, we're able to eliminate the impact of starting wages and determine if job training resulted in increased wage growth for participating workers.

2.2 Exploratory Data Analysis The distribution of change in earnings follows an approximately normal distribution, with a range between -25257 to 60308 dollars and a mean of 2234 dollars. Boxplots were used to assess the relationships between categorical variables and change in wages , while scatter plots and fitted lines were used to visualize possible associations between continuous variables and the response variable.

Change in wages vs Treat When comparing the boxplot of the wage change for workers who received job training to those that didn't, the mean values differ. Compared to workers without job training, workers who received the job training appear to have had a greater increase in income from 1974 to 1978.

Change in wages vs other variables Preliminary data analysis suggests that marriage and age have an effect on the wage changes for workers. Using boxplots to visualize the relationship between wage change and marriage, we observe a clear difference in the average value. This pattern suggests that married workers tend to earn less than unmarried workers. Using scatterplots to examine the association between age and changes in wage, we observe that as age increases wage change decreases. We will explore the significance of the relationships between the response, `inc_78`, and the two predictors, marriage and age, in the model selection process and fitting processes.

Interactions effect

Interaction effects occur when the impact of one variable on the response depends on the value of another variable. This section of the analysis explores the potential impact of demographic groups on the relationship between job training and wage change.

Job training appears to have a substantial effect on the relationship between age and wage change. The difference in slope indicates that workers without job training tend to earn less as they get older. In contrast, workers who received job training see their incomes increase as they age. The significance of this relationship will be explored further in the model selection and fitting process.

There also appear to be interactions between our response and other variables such as no-degree, black, and marriage. The slope between job training and annual wages differs for each of those demographic groups. To draw a solid conclusion on the significance of interaction terms, we will determine their efficiency in the modeling selection process.

3. Model

3.1 Model Selection The model selection is based on two methods: Stepwise AIC Selection and ANOVA F-test.

The final stepwise regression gives us three resulting variables – treatment, age, and marriage. To test the significance of these resulting predictors along with any interaction effects of interest, we incrementally tested each element against the baseline AIC model with an ANOVA F-test.

When comparing the model with interactions between age and treatment and the baseline stepwise regression, we get a p-value of 0.003 and confirm the significance of the interaction term for predicting the changes in wage. Other individual predictors such as nodegree, Black, and Hispanic all perform poorly on the ANOVA test (p-value > 0.05), indicating that they are not significant predictors of wage changes from 1974 and 1978.

3.2 Final Model

$$\widehat{Y}_{Change} = 6072.02 - 4586.31X_{treat:1} - 135.79X_{age} - 1756.52X_{married:1} + 255.88X_{treat1:age}$$

The predictors in our final model for predicting changes in annual earnings from 1974 to 1978 are treatment, married, age, and the interaction effects between age and treatment. All predictors are significant at a significance level of 0.1. The coefficient of determination (R2) of the model is 0.074.

Table 1: Fitting linear model: change ~ treat + age + married + treat:age

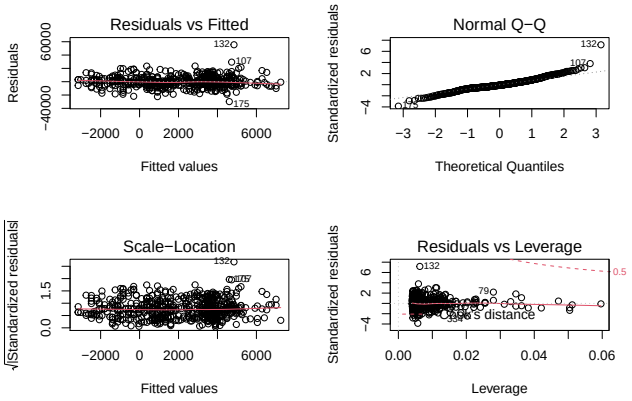
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6072	1044	5.818	9.623e-09
treatTreatment	-4586	2384	-1.924	0.05482
age	-135.8	37.11	-3.659	0.0002754
married1	-1757	715.7	-2.454	0.0144
treatTreatment:age	255.9	87.21	2.934	0.003472

Holding other variables constant, job training is correlated with an average decrease in income change between 1974 and 1978 of $4586, when compared to the baseline of untrained workers. We are 90% confident that the true average decrease in annual earnings is contained in the interval ($659, $8513). However, it must be noted that for workers who receive training, aging is correlated with an increase in income on average while aging is inversely correlated with wages for untrained workers. Compared to untrained workers, for those who receive the training, increasing age by one unit is correlated with a 120 dollar net increase in annual wages.

Holding other variables constant, a one unit increase in age will result in a 135 dollar decrease in annual earnings from 1974 to 1978.

Holding other variables constant, married workers saw a change in wages that was 1756.52 dollars lower than unmarried workers.

3.3 Model Assessment



VIF table for final model

treat	age	married	treat:age
12.21	1.371	1.27	12.08

The residuals vs fitted plot is randomly distributed, which does not lead us to believe there to be any other underlying relationships in the data. As such, the assumptions of independence and equal variance are sufficiently satisfied in our final model. In the QQ plot of residuals, the majority of points fall on the straight line, suggesting there's no violation of the normality assumption. Plotting the residuals vs age, we determined there's no violation of linearity assumption as the points appear to be randomly distributed.

Using Cook's distance, we investigated the data for outliers. As all the Cook's distances were determined to

be far below the 0.5 threshold, we can be confident that our data does not contain any influential outliers. Finally, variance inflation factors were used to confirm that no multicollinearity existed in the chosen model.

3.4 Model Validation

	Final Model	Raw model
Average RMSE	7707	7751

To determine the efficiency of our final model, the k-fold validation and root-mean-square error is used to measure the differences between values predicted by our final model and the value observed. As we compare the current RMSE (7707) to the RMSE of our raw model which includes only main effects of all predictors (7750), we confirmed that the final model is a slight improvement over the original model.

4. Conclusion Consider the effect of treatment alone, there's a negative association between job training and annual wages. On average, trained workers tend to earn 4586 dollars less annually compared to untrained workers. We are 90% confident that the true decrease in income when compared to untrained workers is contained in the range [$659, $8513]. However, the effect of treatment differs as workers age. Trained workers increased their incomes by 120 dollars for every year they aged, compared to untrained ones whose incomes decreased. Marriage also has a significant influence on the change in annual wage. Married workers saw their incomes increase by 1756.52 dollars less than unmarried workers.

5. Limitation There are many limitations of this analysis and the final model in terms of the ability to explain the change in worker wages. First, the sample size is not sufficient enough for exploring interactions between all our predictors, as we lack observations in some levels of education or for certain racial groups. Finally, the low coefficient of determination indicates the model only accounts for 7.4% of the variability in the response. While we could have improved our R2 by including more predictors in our model, we chose to only focus on statistically significant predictors. A model that can more accurately predict changes in worker wages would likely have to include additional data not included in this analysis.

Part II

1. Introduction

There is a common assumption that job training is efficient in boosting the earnings of disadvantaged workers. However, in the 1970s, several experiments were conducted to explore the real impact of job training on wages, including the National Supported Work (NSW) Demonstration[1]. In the experiment, eligible workers were assigned to receive job training, and their incomes in 1973, 1974, and 1977 were recorded accordingly.

This analysis's primary goal is to assess whether receiving job training has a significant result on a worker's probability of obtaining a non-zero wage. Other considerations include whether this effect differs by demographic group (i.e., age, education, racial identity). We are also interested in exploring any other associations with positive income.

2. Data

2.1 Data Preprocessing The original data set is accessed via the NSW Demonstration experiment with 614 non-empty observations. The response variable from this dataset we chose to observe is the participant reported income in 1978, `re_78`. However, since this analysis features the probability of having positive earnings, we created a factorized, binary variable `inc78_factor` - where "0" represents a worker with zero income, and "1" indicates a worker with any positive income. This variable, `inc78_factor`, is our response

variable for the analysis. The cleaned dataset included six categorical variables – marriage, high school degree, Black, Hispanic, education, treatment– and one discrete variable, age.

2.2 Exploratory Data Analysis To get a general understanding of the baseline probabilities of each outcome, and to ensure that sufficient observations for both levels are included in the dataset, we constructed a table of the response variable in order to visualize its distribution. Among the 614 participants observed in the study, the probability of having a non-zero income in 1978 is ~76%.

Conditional possibility tables assess the relationship between categorical variables and non-zero wages while boxplots are generated to visualize the possible association between numeric predictors and non-zero wages.

Response variable vs treat There could exist a negative correlation between treatment and the possibility of having non-zero income. Given the condition where a worker received job training, the conditional probability of getting positive pay decreases slightly compared to those who did not. However, further exploration of this relationship using a chi-squared test (p-value = 0.77) suggests that the two variables are independent of each other. This relationship with treat, and any interactions between treat and other predictors will need to be further examined in the final model.

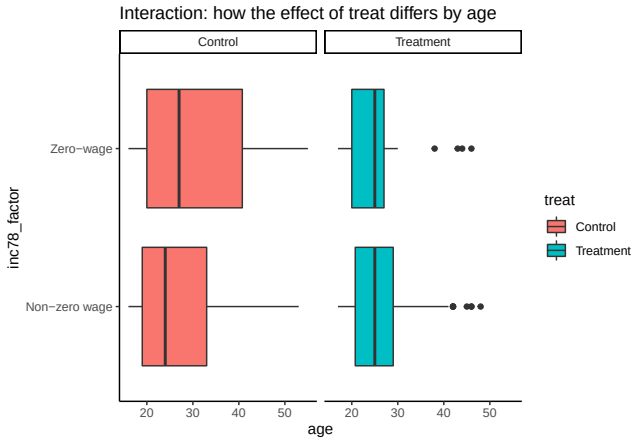
Table for `inc78_factor` vs `treat`

	Control	Treatment
Non-zero wage	0.7716	0.7568
Zero-wage	0.2284	0.2432

Response variable vs other predictors Observing other categorical variables, there appears to be an association between Black and non-zero income. A conditional probability table shows that Black participants have a lower probability of having non-zero income when compared to workers of other racial groups. Preliminary data exploration suggests a difference in the response between non-Hispanic and Hispanic participants, however there are insufficient Hispanic participants to appropriately evaluate this relationship. Changes in marriage and high school degree predictors do not appear to have an effect on the probability of workers earning a positive income.

For our continuous variable, age, boxplots were used to analyze its relationships with the response and to look for potential interactions with other predictors. For participants that reported a positive income, the average worker age appears higher than the zero income group. Further exploration of this relationship using a chi-square test (p-value = 0.002) suggests that the two variables are not independent of each other. At this stage, the probability of having non-zero income appears to decrease as the worker ages, however, considering the small sample size in individual age groups, the significance of the effect age has on the response will need to be re-examined when fitting the model.

Interaction effect Interaction effects occur when the impact of one variable on the response depends on the value of another variable. Conditional possibility tables and faceted boxplots are used to test the interaction effect between variables.



The relationship between non-zero income and age is different for participants that did and did not receive training, indicating a potential interaction effect between treatment(training) and age. For trained workers, positive income probability appears to increase as the worker ages, while for untrained workers age appears to be correlated with a decrease in the probability of earning a positive income.

The association between inc78_factor and age changes significantly when comparing participants with and without high school degrees, indicating a potential interaction between age and nodegree. Positive income probability appears to stay the same, or even potentially increase as a worker with a high school degree grows older, while for non high school graduates the probability of earning a positive income decreases as they age.

To come to an evidence-based conclusion on the interaction terms we will evaluate their significance in the modeling selection process.

3. Model

3.1 Model Selection The model selection followed two methodologies: AIC Stepwise Selection and ANOVA chi-square test.

At the end of the process, the resulting variables are age and black .To test the significance of these resulting predictors along with any interaction effects of interest, we incrementally tested each element against the baseline AIC model with an ANOVA chi-square test.

The inclusion of the interaction of age and treatment is a significant predictor in our model (p-value = 0.038). We will include the interaction in our final model predicting positive wage probability.

Even though the interaction effect between age and high school degree appeared significant in EDA, assessment of the interaction found the term to be negligible. Other individual predictors such as nodegree, married and Hispanic have poor performance on ANOVA test (p-value > 0.05), indicating they are not significant predictors of positive wage probability.

3.2 Final Model At a 0.1 significance level, the four significant predictors in our final model are black, treatment, age and the interaction term between age and treatment. The residual deviance of the final model is 643.51 which is an improvement from the null deviance (666.5). This provides evidence that the model with the predictors is reasonable for predicting the probability of earning non-zero wage.

$$log(\frac{\pi_i}{1-\pi_i}) = 2.558 - 0.753X_{black:1} - 1.289X_{treat:1} - 0.04X_{age} + 0.06X_{treat1:age}$$

Table 5: Fitting generalized (binomial/logit) linear model: inc78_factor ~ black + treat * age

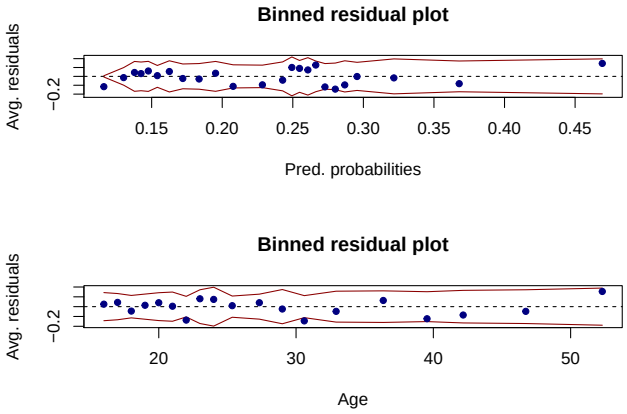
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.558	0.3473	-7.365	1.772e-13
black1	0.753	0.2462	3.059	0.002222
treatTreatment	1.289	0.7456	1.728	0.0839
age	0.04008	0.01036	3.867	0.00011
treatTreatment:age	-0.06038	0.02722	-2.218	0.02655

Keeping other variables constant, compared to people without job training, the average odds of having a non-zero pay for trained workers is 72% lower. We are 90% confident that the true decrease in odds is between 7% and 92%. However, it must be noted that the effect of treatment on positive income probability differs according to age. For trained workers, a one unit increase in age is correlated with a 2.3% net increase in the odds of earning a non-zero wage, when compared to untrained workers.

Holding other variables constant, a one unit increase in age will result in a 3.9% decrease in the average odds of earning a non-zero wage.

Holding other variables constant, the odds of earning a non-zero wage is 52.9% for Black workers, compared to the baseline of white workers.

3.3 Model Assessment



To assess our final model we analyzed the binned residual plots between the raw residuals for fitted logistic regression vs each predictor as well as the residuals vs predicted probability. Ninety-five percent of the observations lay within the plot boundaries, and the plot appears to be sufficiently random. Model assumptions do not appear to be violated.

VIF table for final model

black	treat	age	treat:age
1.586	12.73	1.245	12.07

Finally, variance inflation factors were used to explore any potential multicollinearity effects between predictors in the chosen model. As all the VIFs were confirmed to be below 5 except the interaction terms, we are confident that there are no significant multicollinearity effects in the model.

3.4 Model Validation When using our model for predictions with a decision threshold of 0.5, the accuracy of the final model is 0.77, model sensitivity is 0.99, and model specificity is 0.04.

To further assess the performance of the classifier, we generated a ROC curve by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis). The highest AUC we can observe is 0.617.

ROC curve: final model vs model with all predictors According to the ROC curve, at optimal decision threshold, sensitivity is 0.650 and 1-specificity is 0.569, with AUC of 0.617. Sensitivity on the y axis measures the true positive rate and among all positive cases the final model classified, 65 percent of them are indeed non-zero wages. Meanwhile, 1-Specificity on the x axis refers to false negatives and for all negative cases the final model reported, 57 percent of them are falsely classified as zero wages. Compared to AUC of 0.5 (no diagnostic ability), AUC of 0.617 indicates that the final model has some diagnostic ability for predicting non-zero wages but the performance is not ideal.

4. Conclusion

Considering the effect of job training alone, the average odds of earning a non-zero pay for trained workers is 72% lower when compared to untrained workers. We are 90% confident that the actual decrease is contained in the interval (7%, 92%). However, the effect of job training differs across ages. Compared to untrained workers, workers who received training tend to have a 2.3% net increase in the odds of earning a non-zero wage as their age increases by one-unit.

Besides treatment and age, Black is also an influential factor for predicting the probability of earning a non-zero wage. Compared to White workers, the odds of earning a non-zero wage is 52.9% lower for Black workers, keeping other variables constant.

5. Limitation

According to the result of ROC, the model does not perform well in indicating zero earnings among workers. An accuracy of 77% also suggests limitations of the model in explaining the response variable. As we mentioned earlier, the sample size is not sufficient enough for exploring the real effect of each variable across all levels, especially the interaction effect between predictors. Most importantly, we fail to consider economic factors that might lead to unemployment. Other confounding factors such as the industry, employment status, hourly payment are missing in the observation.

Appendix