Team Green: Clarissa, Tigran, Abhijith, Marlyne, Shufan

Summary

This analysis investigated the effect of job training for disadvantaged workers on their wages based on the data from the National Supported Work (NSW) Demonstration experiment conducted in the mid-1970s. We also looked at the impact of a range of demographic factors. In Part I, our analysis showed that job training increased the difference in wage before and after the training by $2572.97 for unmarried men of 27 years old (average age of the study). In Part II, it was found that for workers of average age who are not black but did complete the training, the odds ratio of having a non-zero wage (or being employed) are 4.55 compared to non-trained.

1 Part I

1.1 Introduction

During the mid-1970s, the Manpower Demonstration Research Corporation (MDRC) operated a temporary employment program for disadvantaged workers across the United States to improve their earnings. The program randomly assigned job training to participants. Participants who received training were in the treatment group, and those who did not in the control group. The program repeated on different workers between 1975 and 1977 (Dehejia and Wahba, 1999). The training guaranteed jobs for the participants in the treatment group for 9 to 18 months and participants who received training were paid for their job. After the training ended, the participants in the treatment group had to find regular employment in the job market like the control group (Lalonde, 1986).

We are interested in evaluating the effect of this training on post-intervention earnings for the male participants in this program - if the workers who received job training tend to earn higher wages than workers who did not. To determine the effect of job training programs accurately, we compared the pre-intervention and post-intervention earnings. Furthermore, it is necessary to account for the variations among individuals in the treatment and control groups. For this reason and also to satisfy other assumptions of the linear regression model selected, we decided to explain the change in earnings for each individual between 1974 (a year before the training started) and 1978 (a year the training program ended) as the response variable.

In this analysis, we fit a multiple linear regression model to estimate the effect of job training on the change in incomes between 1974 and 1978. We also investigate any significant associations between the changes in earnings for male participants and various demographic factors, including race, marital status, years of education, and with or without a high school degree.

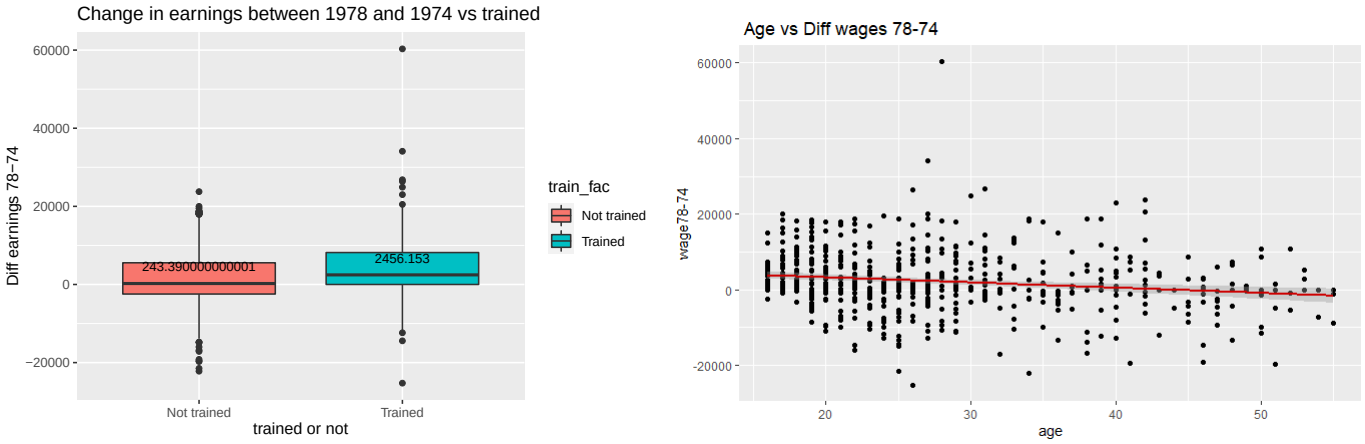
1.2 Data

We look at a subset of the data from the NSW Demonstration containing only male participants. The data includes 614 individuals in total, with 429 in the control group and 185 in the treatment group. Among the 614 individuals, 243 people were identified as Black and 72 as Hispanic, 225 people were married, and 387 did not have a high school degree.

All binary categorical variables are factored - whether a participant is identified as Black (black_fac) or as Hispanic (hispan_fac), whether he was married (married_fac), and if he did not have a high school degree (nodegree_fac). We also centered the age variable at its mean, and named it age_c. No transformation was done on the other numerical variable, education year (educ). We created the response variable re78-74 by subtracting the wage in 1974 (re74) from 1978 (re78).

The contingency table of each categorical variable vs. training shows the sample size for married participants who had training is 1/5th of those who were not married but had training. Furthermore, there is a larger ratio of Black people that were trained compared to non-Black people. The histogram of the response variable, re74-78, shows that its distribution is approximately normally distributed. This confirms the assumption of normality on the response variable.

For each categorical variable, we use side-by-side boxplots to compare the distributions of the response variable for each level of the variable. For the participants who received training, the median change in earnings is $2456, compared to $243 for those who did not receive the training (Fig 1a). This implies that training is a potentially influential predictor on the response variable. The medians of re78-74 are noticeably larger for Hispanic, non-married, non-high school degree participants compared to their respective counterparts. However, the medians of the change in earnings for Black and non-Black participants are approximately the same.



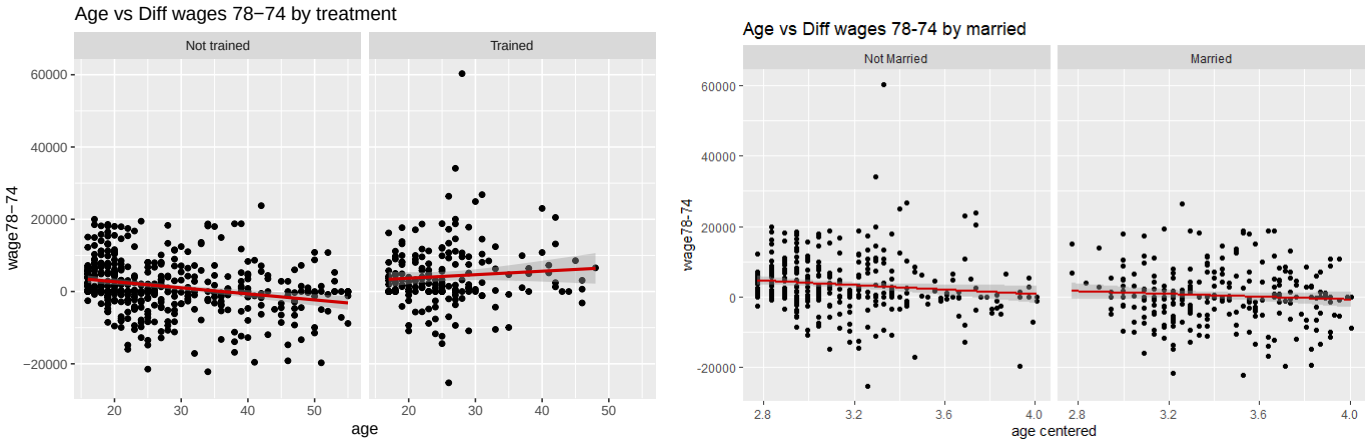
(a) (b)

Figure 1: a) The side-by-side boxplots of the response variable (re78-74) between the treatment and the control groups. b) re78-74 plotted against age

The scatter plot of age vs. the change in wages between 1978 and 1974 (Fig 1b) shows a negative trend between the two. This means that as a person gets older, the wage change is lower. It is the opposite case when looking at education vs. wage difference. People with more education seemed to have a somewhat higher positive change in wages.

A plot of age vs. wage difference split based on training contains two differing slopes. A non-trained participant experiences a negative difference in wages as they get older compared to trained participants who seems to have a more positive change. This suggests that there might be an interaction between training and age. The same could be said for education vs. difference in wages

by training as non-trained seems to have a very flat trend whereas those that had training tend to experience a slightly positive trend. Another interaction we noticed is between age and marriage. The wage difference seems to have a negative trend for non-married individuals as they get older while this trend is neutral for married individuals.



(a) (b)

Figure 2: Comparing the trend of re78-74 VS age a) between the treatment and the control group; b) between married participants to unmarried participants

A plot of age vs difference in wages by Hispanic shows that there might be an interaction between age and the Hispanic factor. The trend for wages tends to go up for non-Hispanic as they get older whereas for Hispanic participants, their difference in wages has a negative trend as they get older (See Fig 8 in Appendix).

1.3 Model

With our main question geared towards understanding whether or not job training leads to change in earnings, the model selected will be a multiple linear regression model with change in earnings between 1978-1974 as a response variable.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n, \tag{1}$$

where y_i is the difference in earnings between 1978-1974 for participant i, the x_i are explanatory variables, and $\beta_i p$, the slope coefficients for each explanatory variable p.

In our first iteration, the model included all main effects where predictors are age centered, married or not, years of education centered, Black or not, Hispanic or not, dropped out of high school or not, received job training or not. Since participants started receiving the training at some point during 1975, we omit using the variable with real annual earnings in 1975 given that there might be some participants who were trained. Based on residual vs. fitted value plots and the normal quantile plot of residual, all model assumptions - linearity, independence, normality, and equal variance - are satisfied with no clear violations.

In our second iteration, the Stepwise AIC and BIC methods are used to select the significant predictors from the model containing all main effects and all possible interactions. Both methods give the same conclusion, and the resulting final model is defined as:

$$\text{re78-re74}_i = \beta_0 + \beta_1 \text{trained}_{i1} + \beta_2 \text{age\_c}_{i2} + \beta_3 \text{married}_{i3} + \beta_4 \text{trained:age\_c}_{i4} + \beta_5 \text{age\_c:married}_{i5} \tag{2}$$

In order to make sure the removed predictors were not significant, we use an F-test to compare this model with different models including either Black_fac, black_fac:trained, Hispan_fac, Hipsan_fac:trained. In all cases, the F-tests suggests a high p-value and we fail to reject the null hypothesis. Thus, we are confident that Eq 2 is final model with the most important predictors.

From here, we check if our model assumptions are satisfied. Linearity is checked by looking at residuals against fitted values (Fig 3). The plot shows residuals are randomly distributed but clustered around two values, 0 and 4000. This means that our model is not able to explain this dichotomous behaviour, but overall, the linearity assumption is valid. It is hard to check for the linearity assumption for age and education variables since they are discrete variables. The normal quantile plot of residual confirms the assumptions on normality, equal variance, and independence. On the normal quantile plot of residuals, even if some points on the tails are not on the 45-degree line, we can still say that there is no clear violation of the normality assumption.

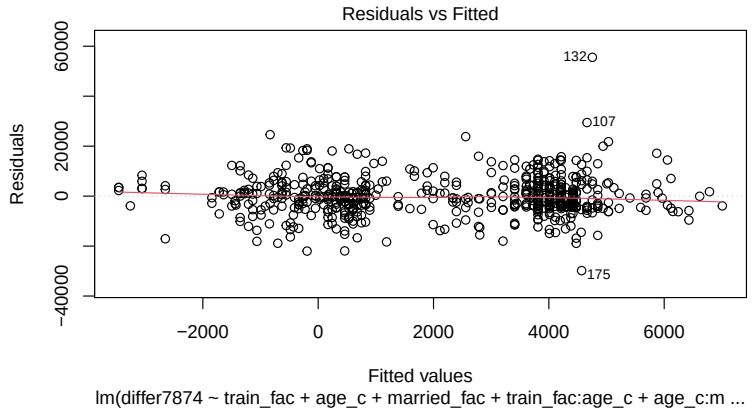


Figure 3: The distribution of a) age, b) re74, for male participants with zero wages in 1978 and those without.

Although there are 3 high leverage points, none of them are influential because they all have Cook's distance smaller than 0.5. The model results stay relatively the same when we take those points out. For that reasons, we decide to keep them in the data. From the VIF metric, no collinearity is found since the values of VIF are less than 10 for all numerical variables.

The summary table for the final model is in Table 1. The value for all coefficients, except the one for the interaction between married and age, have significant p-values. The coefficient on training is significantly not zero, and its 95% confidence interval is $1143 \sim 4002.63$, suggesting the effect on training is significant. The estimated coefficient means keeping the married variable fixed, 27.36 year old males who were trained, had on average \$2572.97 more change in earnings between 1974 and 1978 than the 27.36 year old males who were not trained.

The intercept means that the average change in earnings between 1978 and 1974 for 27.36 year old males who were not trained and unmarried is 1074.79 dollars (95% CI between 1074.8 and 3169.76). The model suggests that marital status and age are influential demographic factors on the response variable. Moreover, as expected from our EDA, the coefficient on the interaction between age and training is significant. The interpretation on their coefficients are the following. The coefficient on age_c means that while keeping other variables fixed, as an untrained man gets older by one year, he gets \$201.94 less change in earnings between 1974 and 1978 with 95% CI (-303 to -100.28). Keeping other variables fixed, married males get on average \$1833.72 less change in earnings between 1974 and 1978 compared to unmarried ones with 95% CI (-3239.05 and -428.4). For the interaction between age

	Dependent variable:
	differ7874
treat_facTreated	2,572.975*** (727.980)*** p = 0.0005
age_c	−201.944*** (51.770)*** p = 0.0002
married_facMarried	−1,833.723** (715.589)** p = 0.011
treat_facTreated:age_c	294.990*** (89.632)*** p = 0.002
age_c:married_facMarried	129.441* (70.764)* p = 0.068
Constant	2,122.276*** (533.375)*** p = 0.0001
Observations	614
R^2	0.079
Adjusted R^2	0.072
Residual Std. Error	7,740.619 (df = 608)
F Statistic	10.452*** (df = 5; 608)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1: Model summary for Part I

and training, keeping other variables fixed, the overall effect of age on change in earnings for males who were trained is \$294.99-\$201.94=\$93.05. We can also say that the additional effect of increasing age by 1 on change in earnings for males who were trained compared to those who were not trained is \$294.99 (95% CI 118.96-471.02).

1.4 Conclusion

The change in earnings before and after job training was examined and modeled to determine if workers who receive job training tend to earn higher total wages than workers who do not receive job training. We decided to model the wage difference instead of the total wages mainly due to the fact that participants in the training group had different baseline earnings among themselves and in comparison with the participants in the control group. Additionally, a model built for the total wages in 1978 would not have satisfied some of the assumptions of a linear regression model given a high proportion of workers had income equal to 0. Our analysis showed that training does have a positive effect on the change in earnings of men who participated in training compared to those who did not. On average, the change in earnings for trained men is \$2572.97 higher. The difference in the change in earnings of trained and untrained men ranges from \$1,143 to \$4,002.63 with a 95% confidence interval. This analysis also shows that the effect of increase in age on the change in earnings is different for men, with the baseline age, who received training than those who did not. For each additional year of age, if the man is not trained, the change in earnings decreases by \$202 dollars, but for the trained man it increases by \$93. Finally, the model also suggests that marriage negatively effects the change in wage. On average, married and untrained men have \$1833.72 less change in wages than those who were unmarried and untrained.

With our model, since we have a very low R-squared value, we might want additional variables from the study to better explain variance in income. Another important limitation is to note that it might take participants a longer time to realize any gains from training and might not be reflected in the 1978 income data. We also had some concerns about data size issues, particularly with race and marital status as the ratio of observations between categories were uneven.

2 Part II

2.1 Introduction

The background of this analysis is the same as in Part I. In Part I we noted that the residuals for our difference in earnings variable (from 1974 to 1978) were clustered around two values, one around 0 and the other at approximately 4000. A reason for this behavior could be this variable is acting as a binary variable, with continuous fitted values off by either 0 or a positive amount. That could explain that many participants were unemployed in both 1974 and 1978, had a decrease in earnings, or became unemployed in 1978 but not before. To distinguish these scenarios, we look at whether a participant was employed in 1978. This analysis investigates if the NSW Demonstration data provides evidence that workers who receive job training tend to be more likely to have employment than workers who did not. The goal is to quantify how job training affects the odds of having non-zero earnings in 1978 and identify other significant influencing demographic factors (same as Part I). Since the employment status in 1978 is likely to be dependent on the earnings in 1974, earnings in 1974 are also included as a potential predictor variable. We apply a multiple variable logistic regression model on the odds of having non-zero wages and the categorical and numerical predictors factors mentioned in Part I.

2.2 Data

This analysis is based on the same dataset from part one where the new response variable is the binary categorical variable re78nonZero. For this response variable we are looking at whether or not workers had nonzero wages in 1978. 471 out of the 614 participants have non-zero wages in 1978, and among the 471 male participants with non-zero wages, 331 of them did not receive training while 140 of them received training. Age and earnings in 1974 are both centered on their means respectively to avoid possible multicollinearity in the logistic regression model. Centering these two numerical variables also makes interpreting the coefficients in the logistic regression model in the next section easier.

For the exploratory data analysis (EDA), we first compare the conditional probability of having non-zero wages in 1978 between the treatment and the control groups with a chi-square test (Table 2a). The chi-square test gives a p-value of $0.77 > 0.05$, suggesting there is no significant difference between the odds of having non-zero wages in 1978 between participants who received training and those who did not. Applying the same approach on Black_fac (Table 2b) shows the conditional probability of being employed differs significantly between Black and non-Black participants ($p = 0.02$). Chi-square tests and comparing conditional probabilities do not suggest other demographic variables, Hispan_fac, nodegree_fac, married are influential factors on the odds of nonzero wage in 1978.

	Not Trained	Trained
zero	0.228	0.243
Non-zero	0.772	0.756

(a)

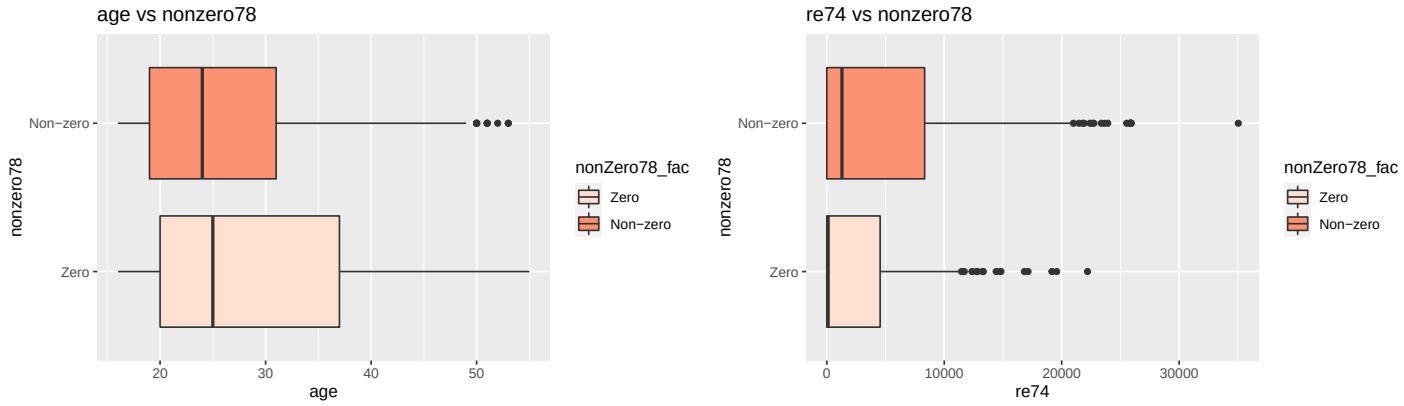
	Not Black	Black
zero	0.200	0.284
Non-zero	0.800	0.716

(b)

Table 2: Conditional probability of having zero wages in 1978 for participants 1) who received training and those who did not, 2) identified as Black and not Black.

For the three numerical variables, age_c, re74_c and educ, the side-by-side boxplots comparing their distributions between participants with zero wage in 1978 and those without shows the medians of age_c and re74_c are different between the two groups (Fig 4a and 4b). On the binned average

odds against age_c plot (Fig 5a), the binned average odds decrease consistently at larger values of ages. These results suggest that age_c and re74_c are likely to have associations with the odds of being employed in 1978. Moreover, our EDA implies that the association between age and the odds of nonzero wage in 1978 is different between the treatment and the control groups. Fig 5b shows that for the participants who did not receive training, the median age of those who were unemployed in 1978 is lower than the median age of those who were employed. However, for the treatment group, the medians of age do not differ much between the participants who had zero wage in 1978 and those who did not. As for the interaction between re74 and training, a similar side-by-side boxplot to Fig 5b shows the medians of re74 for participants who were employed in 1978 and those unemployed are both about 0 in the treatment group, while the medians of re74 are higher for participants who were employed in 1978 in the control group (See Fig 9 in Appendix). This suggests that the interaction between re74 and training might be significant. Besides these two interactions, no other interaction between a categorical variable and a numerical variable have noticeable influence on the odds of a participant having non-zero earnings in 1978.



(a) (b)

Figure 4: The distribution of a) age, b) re74, for male participants with zero wages in 1978 and those without.

2.3 Model

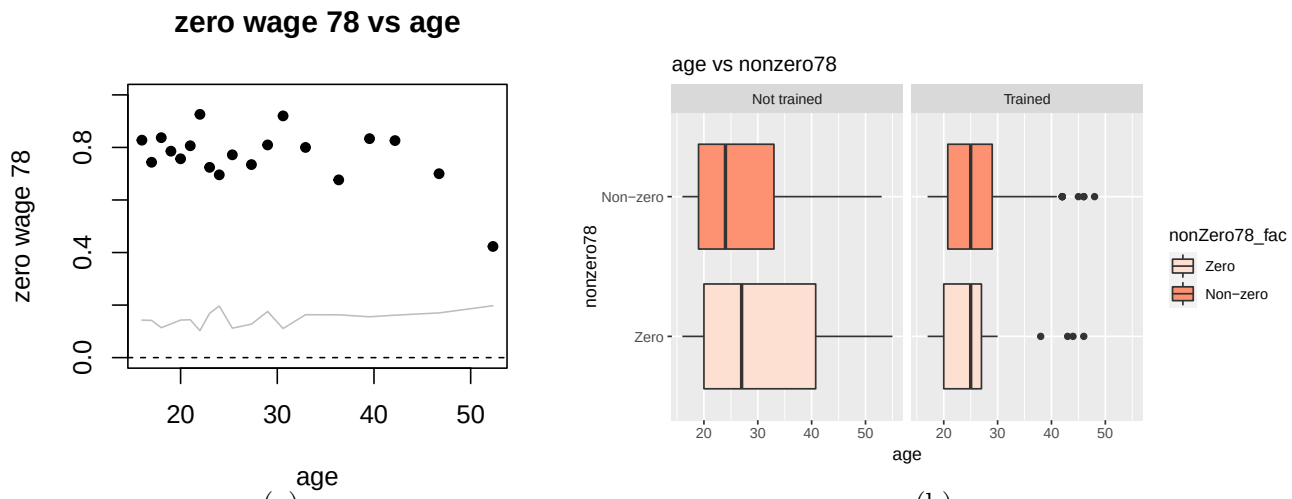
Since our main question gears towards understanding whether a worker who receives job training is likely to have non-zero wages in 1978 than a worker who did not receive job training, the model selected will be a logistic regression model with the binary variable nonzero78 as a response variable.

$$\log(\frac{\pi_i}{(1-\pi_i}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}; y_i|x_i \sim \text{Bernoulli}(\pi_i), \tag{3}$$

where y_i is whether or not a worker has nonzero wages in 1978 i, x_i are predictor variables.

In our first model iteration, we include our main effects where discrete variables are centered. The binned residual plot with overall fitted values has all points within the 95% bin boundary, and residuals are mostly random. However, when we look at the binned plot with residuals versus age, even if most points are within the 95% bin boundary, there is a downward trend starting at age 40 6a. The binnedplot residuals versus education shows no clear trend and all points are within the 95% bin boundary.

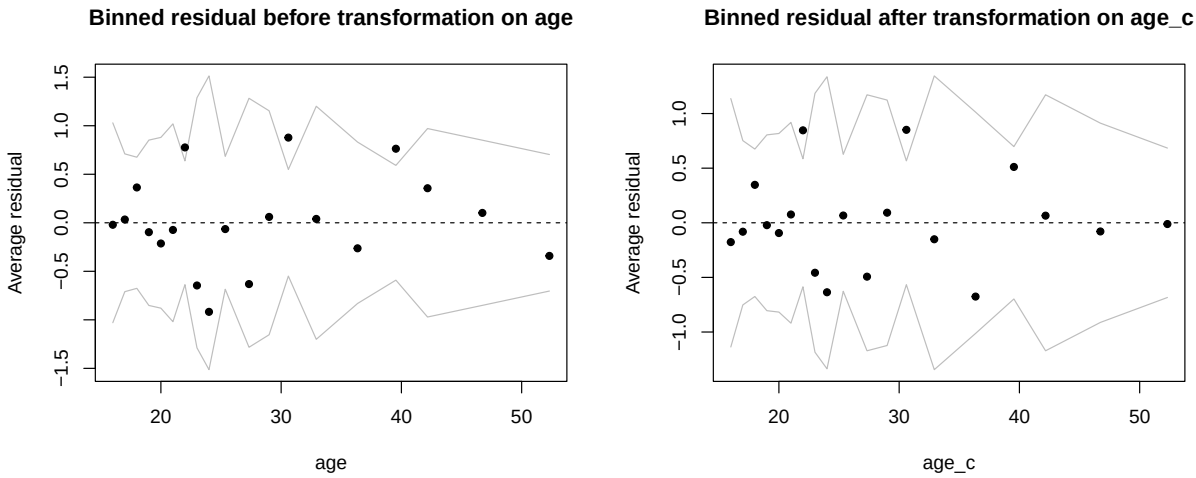
In our second model iteration, we add interactions we think will be significant according to our EDA and use the AIC Stepwise model to select the most important variables. For the purpose of



(a) (b)

Figure 5: a) The binned average odds of having non-zero wage VS age b) the distributions of age for male participants who had zero wages in 1978 and those didn't compared between the treatment (Trained) and the control (Not Trained) groups

our analysis, AIC was picked over BIC since AIC keeps more interactions we are interested in for exploration. Since the new model assessment on age still displays a trend on the average residual plot versus age, we explore different polynomial transformations on this predictor. We start with a second order polynomial transformation but there is no clear improvement on the binned plot. Afterwards, we continue with a third degree polynomial transformation and decide to use it in our final model since it improves our binned plot residual (Fig 6a).



(a) (b)

Figure 6: Binned residual plotted against age. a) before polynomial transformation, a clear downward trend at larger value of age. b) after polynomial transformation, the distribution of the binned residual is approximately random

We also explored using factored education in 5 levels(less than middle school, middle school, high

school, college, higher than college), similar to Part I, but the deviance test confirmed that education, whether factored or not, is not significant. We also created a new binary variable whether or not a worker had positive wages in 1974 which was not picked up by the Stepwise selection using AIC.

Final model results after using AIC and polynomial transformation are:

$$\begin{aligned} \text{Nonzero78}_i|x_i &\sim \text{Bernoulli}(\pi_i) \\ log(\frac{\pi_i}{1-\pi_i}) &= \beta_0 + \beta_1 \text{age\_c}_i + \beta_2 \text{age\_c}_i^2 + \beta_3 \text{age\_c}_i^3 \\ &\quad + \beta_4 \text{re74\_c}_i + \beta_5 \text{trained}_i + \beta_6 \text{black}_i \\ &\quad + \beta_7 \text{trained:black}_i + \beta_8 \text{trained:re74\_c}_i + \beta_9 \text{trained:age\_c}_i \end{aligned} \tag{4}$$

After checking for multicollinearity and outliers, the following summary is generated from our final model (Table 3). From the table, the significant coefficient estimates are: the intercept, trained, third degree order of age, annual earnings in 74, and the interaction between age and trained. This means that, for 27 year old workers who are not black and not trained, the odds of having a positive wage is 3.6 with a 95% CI(2.47-5.34).

	Dependent variable:
	nonZero78_fac
train_facTrained	1.514^{**} $(0.758)^{**}$ p = 0.046
age_c	-0.012 (0.021) p = 0.562
I(age_cˆ2)	0.02) p = 0.179
I(age_cˆ3)	-0.0002^{**} $(0.0001)^{**}$ p = 0.034
re74_c	0.0001^{***} $(0.00002)^{***}$ p = 0.001
black_facBlack	-0.415 (0.288) p = 0.149
train_facTrained:age_c	0.056^{*} $(0.029)^{*}$ p = 0.058
train_facTrained:black_facBlack	-1.264 (0.809) p = 0.119
Constant	1.281^{***} $(0.196)^{***}$ p = 0.000
Observations	614
Log Likelihood	-309.605
Akaike Inf. Crit.	637.210
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: Model result summary for Part II

Keeping all other variables constant, for every 1 dollar increase in 1974 annual salary, the odds of getting positive wages in 1978 for 27 year old workers who did not get trained is 1.00 with 95% CI(1.00-1.00). This shows that even if the coefficient estimate is statistically significant based on its p-value, its effect on the response variable is minimal.

On the other hand, keeping all other variables constant, the odds ratio of getting a positive wage for a 27 year old who got trained compared to another 27 year old who did not get trained is 4.55 with 95% CI(1.28-29.09). It is interesting to note the huge range in confidence interval which implies the different types of workers who got trained. Some people have better odds of getting positive wages than others. For the interaction between age and trained, we can say that keeping other variables constant, as you get older by one year, the odds of getting positive wages for people who got trained compared to the ones who didn't get trained, increases by 5.77%. Since we have a third degree

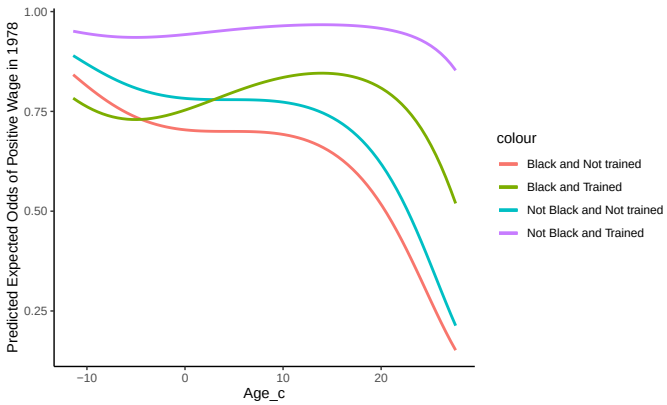


Figure 7: Predicted odds of having non-zero wages as a function of age for different demographic groups to illustrate the effect o the interaction terms and the higher order terms of age in Eq 4.

polynomial coefficient for age, it is better to look at (Fig 7) where as age increases, the predicted expected odds of positive wages decrease differently based on if a participant is Black or received training. It is interesting to note that the drop in expected odds happens later for trained workers compared to non trained ones. On top of that, Black workers who were trained start at the lowest level of expected odds but as they age, they surpass anyone who was not trained.

Using the optimized cutoff value from the ROC curve, the confusion matrix gives an accuracy of only 48.05%, sensitivity is 38.22%, and specificity is 80.42%. This means that given that someone has positive wages, our model correctly classifies it 38.22% of the time which is not the best. On the other hand, given that someone had non positive wages, our model correctly classifies it 80.42% of the time. On top of that the ROC curve gives a 65.5% AUC score which is not ideal.

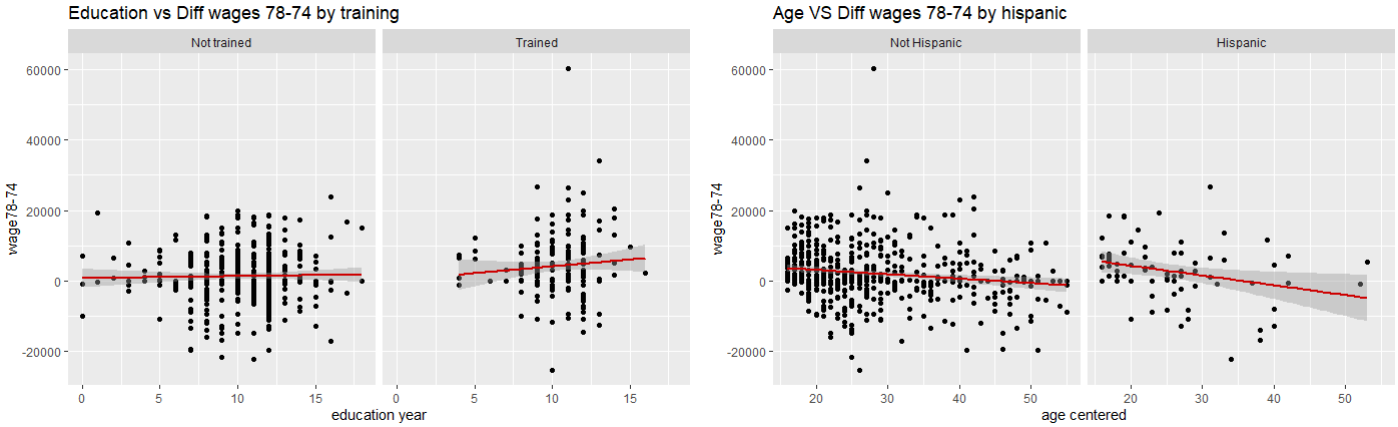
2.4 Conclusion

A logistic regression model with multiple predictors was used to determine whether training is an important factor in having positive wages in 1978 or not, and to look for variables that have a significant effect on it. The regression result shows the significance of the predictors at the 5% significance level except Black male variable and its interaction with training. We concluded that for 27 year old males, training increases the odds ratio of having positive wages 4.55 times compared to ones with no training. Although we did not find any significant effect in demographics such race (Black or Hispanic), both age and its interaction with training were statistically significant. In addition, the impact of 1974 wages on the response variable is minimal.

The predictive capability of the model is not ideal. It would not be practical to use this model to classify workers with 0 or non-zero wages due to low accuracy of only 48.05%, sensitivity is 38.22%. A cubed variable (age) in the model creates difficulties in interpreting the coefficients. A graph (Figure 8) must be used to understand the relationship of age and the odds of earning non-zero wage in 1978. As explained in the Limitations section in Part I. The timing of the training could be an important aspect that the data and model do not capture. The same doubts with race and marital status as the ratio of observations between categories are very uneven.

3 Appendix

3.1 Part 1 Appendix



(a) (b)

Figure 8: Exploring the interaction between a) education and training, b)education and Hispanic.

3.2 Part 2 Appendix

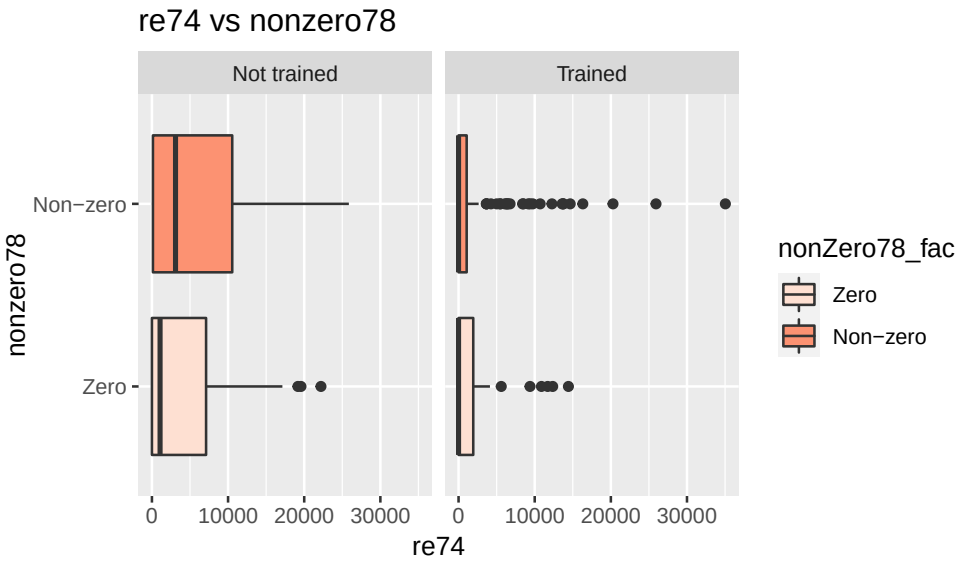


Figure 9: Exploring the interaction between training and re74 on the odds of having non-zero wage in 1978. The relative difference in the median of re74 between participants on non-zero wage and zero wage in 1978 is different for the treatment and the control group.