Does Job Training Improve Wages?

Team: Erika Fox (Presenter), Minjung Lee (Coordinator), Preet Khowaja (Writer),
Raza Lamb (Programmer), Robert Wan (Checker)

10/7/2021

Summary

This analysis uses linear and logistic regressions to study the effect of a job training program on the wages of male workers in 1978. Findings from this report suggest that while training alone did not have a statistically significant affect on wages, its effect was dependent on demographic factors such as age and income before the program.

Introduction

Since the peak of the industrial revolution, the demand for unskilled labor has been declining. Some of the biggest culprits are globalization and advances in technology that reduce the need for farm and manufacturing labor. Furthermore, with the impending advances in AI, there is reason to believe this trend will continue, if not escalate. Thus, the efficacy of job training is a continually highly relevant topic as policies are developed to advance the workforce.

In this analysis, we investigated the relationship between job training for disadvantaged workers and their wages after the training has concluded using data from an randomized experiment conducted at the National Supported Work (NSW) Demonstration. In the NSW Demonstration experiment, eligible workers were randomly assigned to receive or not receive vocational training between March 1975 and July 1977. For this study, we used a subset of the NSW Demonstration data. The test group of this study is a group of selected males who were received job training and for whom the salary in 1974 is available. The control group is a group of males whose income in 1975 was below the poverty line. This analysis is split into two parts: Part I uses a linear regression model with multiple predictors and Part II uses a logistic regression model with multiple predictors.

Part I

We fitted a linear regression model to investigate whether workers who received training were likely to have higher wages in 1978 compared to those who did not receive training. We also wanted to answer these questions:

- What is a likely range for the effect of training on income in 1978?
- Is there evidence that the effect of training differed by demographic groups?
- Are there other interesting associations with wages?

Part 2

We utilized logistic regression to determine whether or not workers who received job training tended to be more likely to have positive wages than those who did not. Other questions of interest include the following:

- What is a likely range for the effect of training on the odds of having a positive income?
- Is there evidence that the effects of training on positive income differed by demographic groups?
- Are there other interesting associations with positive income?

Part I: Linear Regression

Data

The data for this analysis comes from an experiment from the 1970's conducted by The National Supported Work (NSW) Demonstration that evaluated public policy programs. The experimenters sought 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.

Code book:

- *treat*: 1 if the participant received job training, 0 if the participant did not receive job training
- *age*: age in years
- *educ*: years of education
- *black*: 1 if race is black, 0 otherwise
- *hispan*: 1 if Hispanic ethnicity, 0 otherwise
- *married*: 1 if married, 0 otherwise
- *nodegree*: 1 if the participant dropped out of high school, 0 otherwise
- *re*74: real annual earnings in 1974
- *re*75: real annual earnings in 1975
- *re*78: real annual earnings in 1978

For our analysis, we assigned the variables *treat*, *black*, *hispan*, *married*, and *nodegree* as factor variables. The remaining variables were either discrete or continuous variables. We decided to use *re*74 as the baseline income variable for the model based on the findings from the EDA. We did not use the variable *re*75 because the training group was paid in 1975 by the program although the control group was selected based on income in 1975. As such, using *re*75 would make it difficult to measure the true effect of *treat* on wages. For more discussion on this, please see the conclusion.

During the initial data exploration, we noticed several features of the data. First of all, there are relatively few men in the data who are Hispanic (only 72), which will be important when interpreting the model. We also noticed that when compared to workers with positive income in 1974, those who had 0 income in 1974 were more likely to also have 0 income in 1978. To account for this in the model, we created a new dummy variable named *zero*. This variable was coded as 1 if the participant had no income in 1974 and 0 otherwise. Another noticeable feature of the data is that there are 18 men with exactly the same wage in the control group in 1974, which also happens to be the maximum wage for the control group in that year. Upon a brief examination, these men do not appear to be duplicate observations, but this is most likely not a coincidence. For this investigation, we left these observations in the data, although they may deserve a closer look in subsequent research. Finally, we noted that the response variable, *re*78, is not normally distributed, and is in fact heavily skewed towards 0. This may be an indication that the research question is not well suited for linear regression. Figure 1 shows the distribution of *re*78.

Before moving on to fitting a model, we examined relationships between the response variable and potential predictors. We utilized scatter plots to investigate the relationship between the response variable and numerical predictors. For categorical variables, we used box-plots. Based on the plots, there appeared to be a positive relationship between the response and both *educ* and *re*74. The response variable might have had a positive relationship with *age* as well, but this trend was not immediately clear. For the categorical variables, we deemed that nearly all of them appeared to have a relationship with the response variable. On the other hand, the predictor of interest, *treat*, appeared to have a very small impact, and those who received training actually tended to have a lower wage, which was contrary to our expectation.

Finally, we investigated all potential interactions and had several interesting findings that were relevant to model building. There appeared to be an interaction between *treat* and *black*: men who are black appeared to benefit more from the training than men who are not black. There is also an interaction between *treat* and *zero*: among men whose income was positive in 1974, the training group had lower average income, while the trend is reversed for men who had 0 income in 1974. Both of these interactions would be especially important

Figure 1: Histogram of Income in 1978

to investigate further during model building due to their direct relation to the questions of interest. Other interactions that we deemed potentially important were: *married* and *age*, *married* and *re*74, *married* and *hispan*, and *married* and *zero*.

Model

We chose to use *re*78 instead of the difference between *re*78 and *re*74 as the response variable of the linear regression model. Using the latter would imply that the starting income level in 1974 did not matter for predicting wages in 1978. We did not believe that was a reasonable assumption since we observed that workers who received 0 income in 1974 were more likely to also receive 0 income in 1978.

Before constructing the model, we centered all continuous predictor variables around the mean and created them as new variables: *agec*, *educc* and *re*74*c*. This would allow for more meaningful interpretations of the intercept in the final model. When we mention *age*, *educ* and *re*74 in model interpretation sections in this report, we are referring to the centered versions of these predictors.

The final model is as follows:

$$re78_i = \beta_0 + \beta_1 treat_i + \beta_2 re74_i + \beta_3 educc_i + \beta_4 zero_i + \beta_5 black + \beta_6 treat_i : zero_i$$

Model Selection

In order to arrive at the final model above, we started with a null model and a full model. The null model consisted of just one predictor, *treat*. The full model included the effects of *treat*, *re*74, *educc*, *zero*, *black*, and *treat* : *zero*, as well the interaction between *treat* and all the other predictors. We also included the interaction terms *married* : *agec*, *married* : *re*74, *married* : *hispan*, and *married* : *zero* because we saw interesting trends for them in the EDA.

We performed forward, backward, and step-wise selection using both AIC and BIC. All three stepping methods using AIC selected the same model, and the stepping methods using BIC selected a different model. We performed an F-Test on the two models to determine which one to pick. The p-value was significant at the 0.05 level, providing evidence that the predictors in the model selected by AIC should be included. Therefore, we picked the model selected by AIC as the final one (see equation above for a formal representation).

Model Assessment

In order to assess the model, we first looked at the residuals vs. fitted values. We notice that the points have a lower bound. This is because the response variable, $re78$, has a lower bound of 0, which means the residuals would also be bounded and not exactly normally distributed. This raises the question of whether a linear regression is the best model for this study. A better model would be the logistic regression model we would used in Part II. We also investigated potential transformations, including taking the natural log and square root of the response variable. However, they did not have noticeable effects.

Setting aside our discomfort with the bound of the residuals, the model did fairly well on normality and equal variance. It's also reasonable that the observations are independent. We had some concerns about the linearity of the relationships, and we also observed a downward trend in the plot of $re74$ against residuals, as shown in Figure 2.

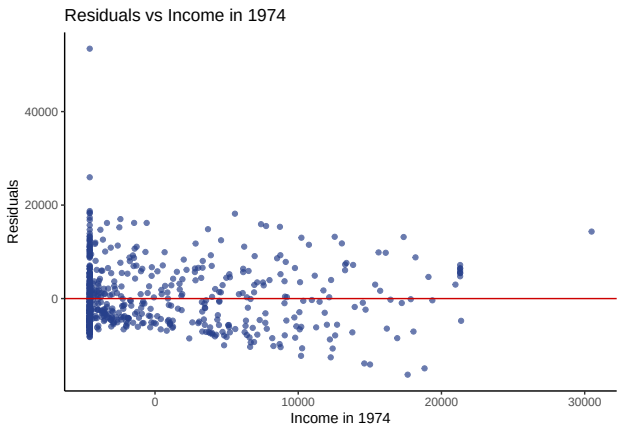


Figure 2: Assessing our model

In an attempt to battle the trends we observed in the residual plots, we added a transformation to the model. We tried to add a square term for $re74$, but it did not improve the residual plot so we decided to forgo a transformation here and went back to the final model shown above. To check for multicollinearity, we calculated the VIF of the predictors. All VIF were well below 5, so we were not concerned about multicollinearity.

Finally, we investigated the model for potential high influence points. We noticed one point with very high leverage, and one outlier whose standardized residual was approximately 8. To determine the effect of those points on the model, we removed both of them and refitted the model. There were no observable effects on model coefficients or accuracy, so we included those points back into the final model.

Interpretation and Results

The results of the final model can be viewed in the table below. The significant predictors were $re74$, $educc$, $black$, and the $treat : zero$ interaction term. We could infer the relationships between those significant predictors and $re78$ based on the model coefficients. In order to determine the ranges of the effects on $re78$, we calculated 95% confidence intervals. All other variables held constant, an increase in income in 1974 by \$1 implied an increase in income in 1978 of between \$0.35 and \$0.56 on average. An increase of 1 year in education was likely to increase income in 1978 by an average of \$313, if all other variables are held constant. If a participant identifies as black, he was likely to earn \$1434 less in 1978 than a participant who does not, provided all other variables are held constant.

From the interaction of $treat : zero$, we inferred that a worker who received training but had 0 income in 1974 was expected to earn approximately \$4,638 more in 1978 and a worker who also received training but had a positive income in 1974, holding all else constant. This interaction term $treat : zero$ was interesting

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6532.7235	436.6906	14.96	0.0000
treattraining	-1187.1874	1123.7577	-1.06	0.2912
re74c	0.4587	0.0529	8.67	0.0000
educc	313.4344	106.9450	2.93	0.0035
zerozero	1017.7921	859.7100	1.18	0.2369
blackblack	-1434.2414	721.6446	-1.99	0.0473
treattraining:zerozero	3668.2062	1349.5765	2.72	0.0068

because it was a significant predictor in the final models for both Part I and Part II, which we will discuss later in the report.

Inferential Questions

Having fit the final model, we could begin to answer the inferential questions. First of all, the model suggested that for a non-Black / non-Hispanic male who had positive income in 1974, the training program actually decreased his wages by an average of $1187.19. The 95% confidence interval for this effect was $(-3394.11,\ 1019.7376405)$. As is visible, this interval included 0, which meant that we could not reject the null-hypothesis that the true effect of training was 0. However, the interaction of $treat : zero$ was significant at the 0.05 level. As discussed in the section above, the effect of training was positive for workers whose income in 1974 was zero.

We also wanted to see whether the effect of training varied by other demographic characteristics. We included all interactions of demographic data in the full model, but none survived selection using AIC or BIC. For our purposes, we concluded that we did not have enough evidence to say that the effect of training varied by demographic characteristics. Finally, we noted that income in 1974 and education had strong predictive power for income in 1978, both of which had a positive relationship.

Part II: Logistic Regression

Data

For Part II of the analysis, we used the same data set as in Part I. Here, we intended to investigate whether this specific job training program increased a worker's odds of having a positive wage in 1978. We used the same variables as in Part I with one exception - in order to answer the main inferential question, we created a new variable named *positive*, which is 1 if the participant had a positive income in 1978 and 0 otherwise. This is the response variable for this analysis. Same as in Part I, we used $re74$ as the baseline income variable. The initial data observations from Part I still hold here. To restate: there are relatively few observations in the data who are hispanic, which will be important when interpreting the model. We also used the variable *zero* in this analysis.

Subsequently, we examined relationships between potential predictors and the response variable. Several variables seem to have a relationship with the response variable, including, *age*, $re74$, *black*, and *zero*. The relationship between education and the resopnse variable is worth mentioning. *Educ* and *nodegree* both reflect the level of education. We verified that *nodegree* did not have an observable relationship with the response variable using a Chi-Squared test, but we did notice a trend in *educ*. Figure 3 shows the mean value of *positive* for each value of *educ*. Interestingly, while there was not a consistent trend, we observed that workers with fewer than 9 years of education had lower mean values of *positive* when compared with workers with more than 9 or more years of education. To further investigate this, we created another dummy variable named *newed*, which is coded as 1 if the worker ha s 9 or more years of education and as 0 otherwise. Using a Chi-Squared test, we confirmed that *newed* was not independent from *positive*, suggesting that we should include it in the model.

Lastly, we investigated potential interaction effects of predictor variables. To examine the interaction of discrete and continuous variables with categorical variables, we created box plots broken out by levels of

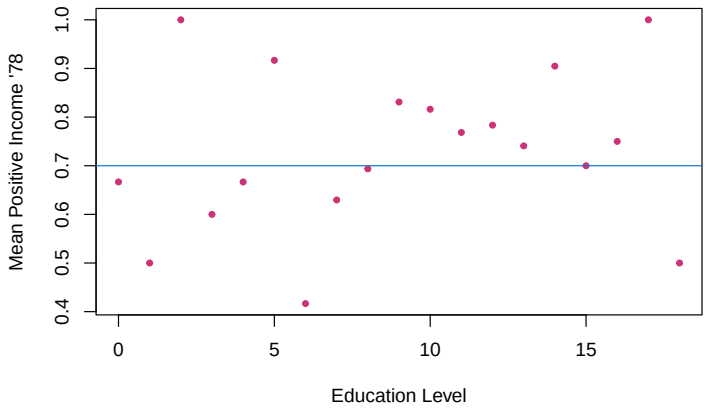


Figure 3: New education variable against positive income-1978

the categorical variables. We used tables for interactions of different categorical variables. We found several interactions terms that we wanted to include in the model, including $treat : re74$, $treat : black$, $treat : age$, $treat : educ$, $treat : married$, $black : re74$, $married : re74$, $educ : black$, and $educ : married$.

Model

The final model we selected is a logistic regression model with multiple predictors, shown below. Similar to the linear regression model of Part I, all continuous variables in this model were centered around the mean.

$$positive\_income\_i|x_i \sim Bernoulli(\pi_i);$$

$$\log(\frac{\pi_i}{1-\pi_1}) = \beta_0 + \beta_1 treat_i + \beta_2 agec_i + \beta_3 agec_i^2 + \beta_4 agec_i^3 + \beta_5 educc_i + \beta_6 black_i + \beta_7 re74c_i + \beta_8 zero\_re74_i +$$

$$\beta_9 nine\_or\_more\_years\_educ_i + \beta_{10} treat_i : agec_i + \beta_{11} treat_i : agec_i^2 + \beta_{12} treat_i : agec_i^3 + \beta_{13} treat_i : zero_i$$

Model Selection

Keeping the research questions of interest in mind, we created a null model with *treat* only. The full model included centered *age*, *educ*, and *re*74, as well as *married*, *hispan*, *black*, *zero*, and *newed*. We also included interaction terms of *treat* with each of the individual predictors because we were interested in exploring whether the relationship between *treat* and *positive* changes by the demographic information that the other variables carry. Lastly, we included interaction terms from the EDA that we found interesting and wanted to investigate further.

We spent a lot of time debating the method to use to select the final model. After constructing the full model, we conducted forward selection, backward selection, and step-wise selection using both AIC and BIC. All three BIC selection processes resulted in the same model, referred to here as *bic_stepwise*. AIC step-wise and AIC forward resulted in the same model (referred to as *aic_stepwise*), and AIC backward resulted in a different model (referred to as *aic_backward*). First, we looked at the ROC curves of each of those models to see which one returned the highest area under the curve. It was clear that the model selected by BIC has lower sensitivity and specificity for nearly the entire path. However, the two AIC models were every similar in both predictive power and area under the curve.

Next, we used Chi-Squared tests to determine which model may be more accurate. The difference between *bic_stepwise* and both AIC-selected models was significant at the 0.05 level. However, the difference between

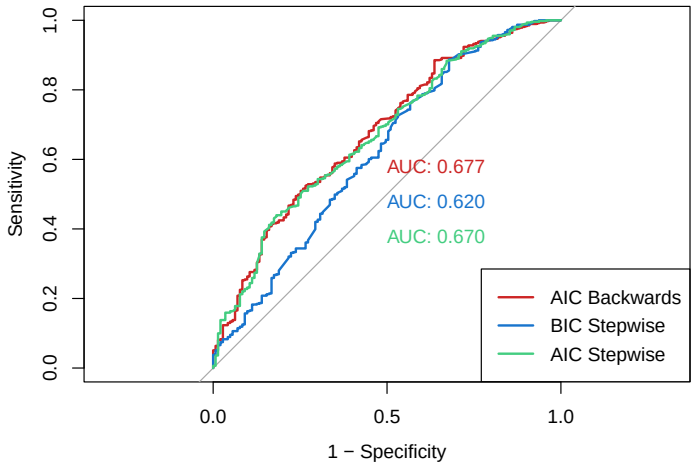


Figure 4: Comparing Model ROC's

the *aic_stepwise* and *aic_backward* was not significant at the 0.05 level. This information, in combination with the ROC curves, suggested that we should remove the BIC-selected model from contention and narrow our attention to the two AIC-selected models.

There were two main lines of thought as we debated over picking *aic_stepwise* or *aic_backward*. While *treat* was significant in *aic_stepwise*, it was not significant in *aic_backward*. However, *aic_backward* kept the interaction term $treat : zero$ and showed it was significant at the 0.05 level. This prompted discussions of which model we should pick and how our interpretations would differ based on the model we picked. If we picked *aic_backward*, we could not say training by itself had a significant effect on the odds of earning a positive wage. On the flip side, it would allow us to interpret the effect of interactions between *treat*, *age*, and *zero*. Given how wage is effected by age and prior income in the real word, we decided it was more reasonable to quantify the effect of training in the context of age and income prior to training. Ultimately, we picked *aic_backward*.

We simplified the *aic_backward* model by removing the predictors *hispan* and $hispan : treat$. A Chi-Squared test revealed that they were not significant predictors and the EDA showed that there were not many observations for men who identified as Hispanic. Upon closer examination of the binned residual plots, we noticed that there were clear trends in the plots for fitted values vs. binned residual and age vs. binned residuals. This current model does not stand true to the assumptions of logistic regressions - we will address the issues in the next section.

Model Transformation and Assessment

In order to address the issue of trends we saw in the residual plots, we explored some transformations of the working model. We found that adding a square centered age term did not visually alter the binned residual plots but further adding a cubic term for centered age significantly improved the binned plots. Since we were more comfortable with these plots, we decided to include both the square centered age and the cubic centered age terms in the model. Figure 5 displays the effect of adding the polynomial terms. Before the addition, there was a decreasing trend in the binned residuals, which was visibly resolved after the additions.

The final model mentioned above includes *agec* squared and *agec* cubed as well as the interaction between *treat* and these two higher-order *agec* terms.

As before, we created binned residual plots against the continuous predictor variables. We checked whether or not the points were mostly within the 95% confidence interval bounds and whether or not they were randomly spread. The model safely satisfied assumptions, and we were comfortable using it for the purpose of this

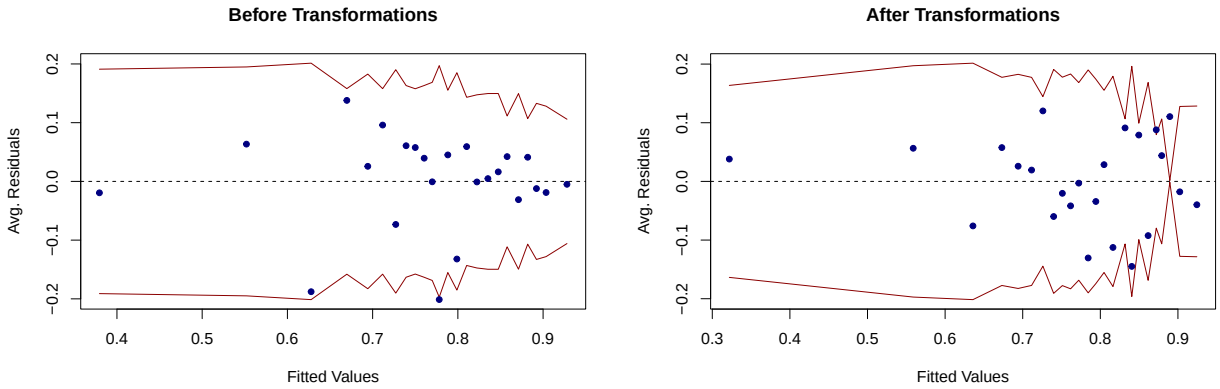


Figure 5: Comparison of Residual Plots

study. There were also no visible high leverage / high influence points or outliers, so we decided proceed with this model.

One caveat for adding higher order terms for age in the model was that we would have high VIF for *agec*, *age*2 and *age*3 since these three predictors were bound to have high multicollinearity. We fitted a model without the higher order terms and confirmed that the model had high VIF only due to the higher-order age predictors. Except for those variables and the interaction terms, no VIF was worryingly high.

Interpretation and Results

The table below shows the output of the final model, in which *age*3, *black*, *re*74*c*, *newed* and *treat* : *zero* are significant predictors. In particular, the odds of having a positive income in 1978 were 0.53 times if the worker is black vs. if he is not, holding all other variables constant. Also, the odds of a positive income in 1978 of a worker with 9 or more years of education were 2.44 times more than a worker with fewer than 9 years of education, all else held constant. *Treat* was not a significant predictor, but *treat* : *zero* was significant.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.7518	0.3345	2.25	0.0246
treattraining	-0.0457	0.4603	-0.10	0.9210
age2	0.0040	0.0026	1.53	0.1269
age3	-0.0003	0.0001	-2.18	0.0296
educ	-0.0776	0.0585	-1.33	0.1844
blackblack	-0.6359	0.2563	-2.48	0.0131
re74c	0.0001	0.0000	2.38	0.0174
zerozero	-0.4976	0.3168	-1.57	0.1163
newed9 or more	0.8903	0.3706	2.40	0.0163
agec	-0.0080	0.0225	-0.35	0.7230
treattraining:agec	0.0526	0.0615	0.85	0.3932
treattraining:zerozero	0.9860	0.4784	2.06	0.0393
treattraining:age2	-0.0010	0.0050	-0.20	0.8418
treattraining:age3	0.0000	0.0004	0.05	0.9567

Because our model includes interaction and polynomial terms, interpretation in words can easily become confusing and messy. To aid interpretation, we created Figure 6. This graph displays how the probability of having a positive income in 1978 changes by *age*, stratified by two categorical variables: *treat* and *zero*. A few insights can be gleaned from the plot. First, the relationship between age and the odds of having a positive income in 1978 was not linear. For younger workers, being relatively older decreased their likelihood

of having a positive income, but this effect leveled out as age increased and eventually reversed for some groups. Then, at approximately 15 years above the average age, there was a steep drop off in the probability of earning a positive wage for all groups. Second, we wanted to note how demographic variables changed this relationship. Whether or not a participant had zero income in 1974 influenced the effect of training. The graph shows that participants with training were more likely to have positive income in 1978 if they had zero income in 1974 compared to someone with positive income in 1974 but still received training. However, the effect was reversed for workers who did not receive training - individuals who were not trained and had zero income in 1974 were less likely to have positive income in 1978 than individuals who were not trained and had positive income in 1974. Finally, the benefit of training also increased with age. The benefit was barely discernible and may have even been negative for the youngest workers in the data set, but it increased from there.

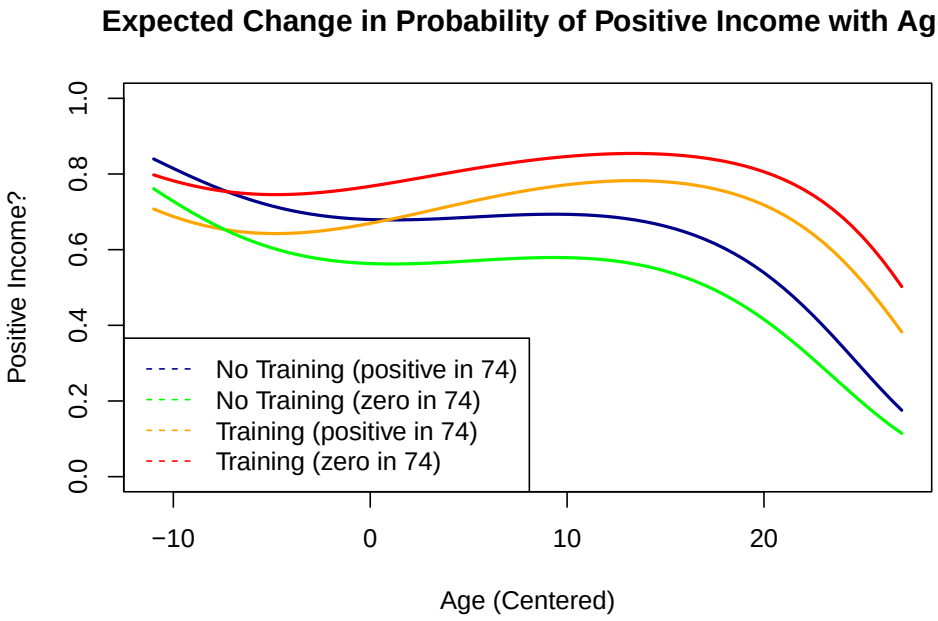


Figure 6: Interpretating Age as a Predictor

Inferential Questions

Having fit the final model, we could begin to answer the inferential questions. The question of whether training by itself effected the odds of having a positive income in 1978 had a straightforward answer. The 95% confidence interval of the effect of training on the odds was (0.3922711, 2.3996102). Since the interval included 1, we failed to reject the null hypothesis that treatment had not significant impact. However, the question of whether training had any effect on the odds could be difficult to answer in words due to the interactions and polynomials. The graph in the previous section does a good job of answering the question. With its help, we came to two main conclusions: (1) the positive effect of training increased with age, and (2) the effect of training was larger for workers with 0 income in 1974.

Except for the two conclusions above, however, interactions between other demographic characteristics and training were excluded in the model selection process. As such, we do not have sufficient evidence to conclude that the effect of training varied by demographic characteristics besides age. On the other hand, there were a few other predictors in the model worth mentioning. First of all, a black worker had lower odds of having a positive income than a non-black worker. Continually, income in 1974 also had a significant effect on the odds of having a positive wage. And finally, the dummy education variable we created *newed* was significant

- this suggested that if workers who entered high school, regardless of whether they completed high school, had higher odds of having non-zero wages in 1978.

Conclusion

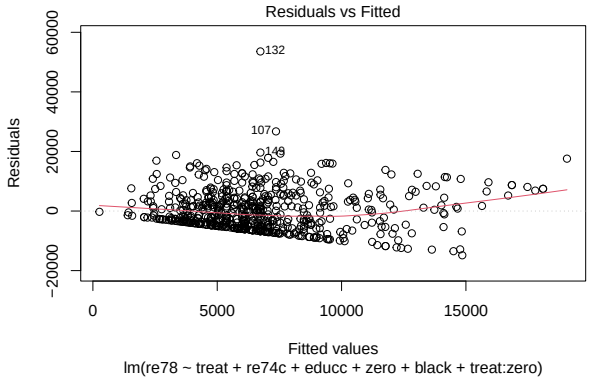
Having conducted both linear and logistic regressions on this data set, we have been able to answer the main inferential questions. Interestingly, in both models, this specific job training did not have a significant effect on income by itself — instead, the effect was conditional on other factors. In the linear regression model, the effect of treatment was dependent on whether or not the worker had zero income in 1974. In the logistic regression, the effect was similarly dependent on the variable *zero*, although it was also dependent on the age of the worker. Besides those, we did not find the effect of training to be dependent on other demographic characteristics.

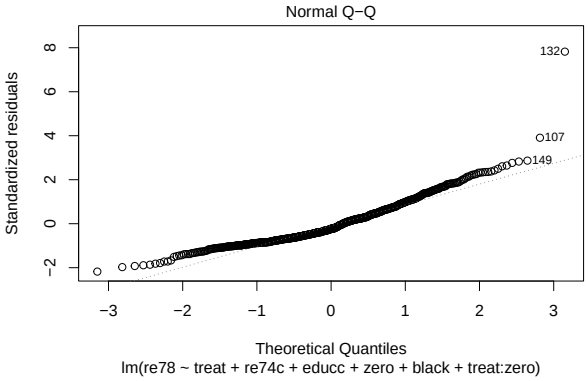
Ultimately, it's unlikely for a specific program to work for all kinds of demographic groups. Both models show that this specific training program was more likely to work for people who had zero income in 1974 than for people who had some positive income in 1974. For the odds of having positive income, we also have evidence that as age increased, the benefit of training increased.

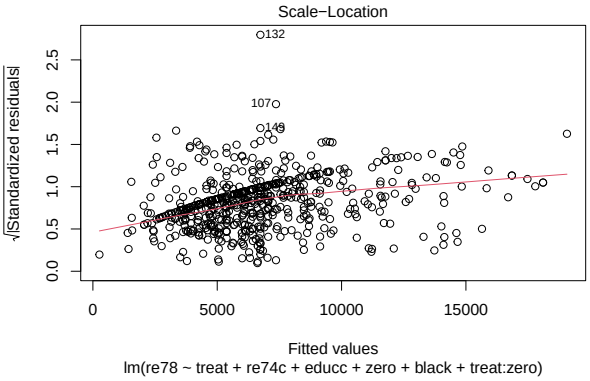
The study was limited by the data and methods we used. The data was insubstantial, and $re75$ was a messy variable. Since workers in the training program were paid basic wages, it was unfair to use $re75$ to predict income in 1978 for tha training group. Yet, we could expect the control group's income in 1975 to affect income in 1978. Data problems extended to existing variables like *hispan*, where we didn't have enough observations to appropriately assess whether Hispanic workers were affected by the training program differently than other racial groups. Furthermore, our interpretations of the coefficients and predictors were specific to the training program in this particular data set. We could not infer too much about training programs at large.

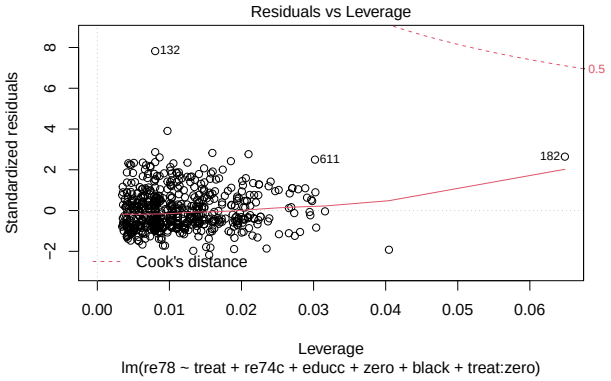
One thing to keep in mind as we read this study is that the training program under scrutiny is from the 1970's. It does not include data on female workers and even if it did, it may not be the best metric to assess training program in modern days. Conclusions we drew from this study need to be understood in the context from which the data was collected.

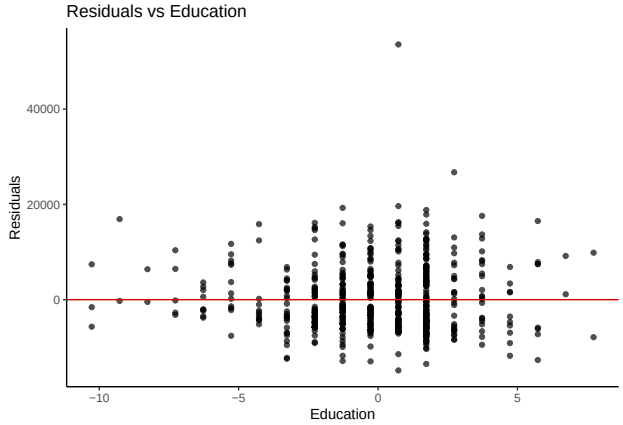
Appendix 1: Final Model Assessment — Linear Regression

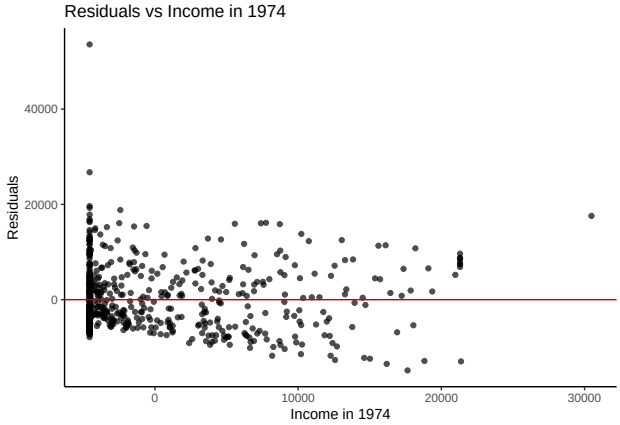












Appendix 2: Final Model Assessment — Logistic Regression