Effects of Job Training on Wages

Aarushi Verma, Deekshita Saikia, Mohammad Anas, Tego Chang, and Sydney Donati-Leach

Introduction

In 1986 Robert J. LaLonde performed a research study to evaluate the econometric evaluations of training programs (National Supported Work Demonstration) on post intervention income level for disadvantaged workers. In our report, we consider a subsection of the data used in the original study to explore questions similar to the ones in the original study:

- Is there evidence that workers who receive job training tend to **earn higher wages** (Part I) and **positive wages** (Part II) than workers who do not receive job training?
- Can the impact of receiving job training on earnings be quantified? What is the likely range of the effect of the treatment?
- Do the effects differ by demographic groups?
- What are the other interesting associations with income?

Part I

Summary

This analysis examines the effect of job training on wages. Exploratory data analysis and step wise selection are performed to determine the regression model. According to our analysis, training has a positive impact on wages for the participants. We also explored interaction effects between age and treatment for the participants. Finally, we found that demographic factors such as age of the participants has a negative impact on wages.

Data

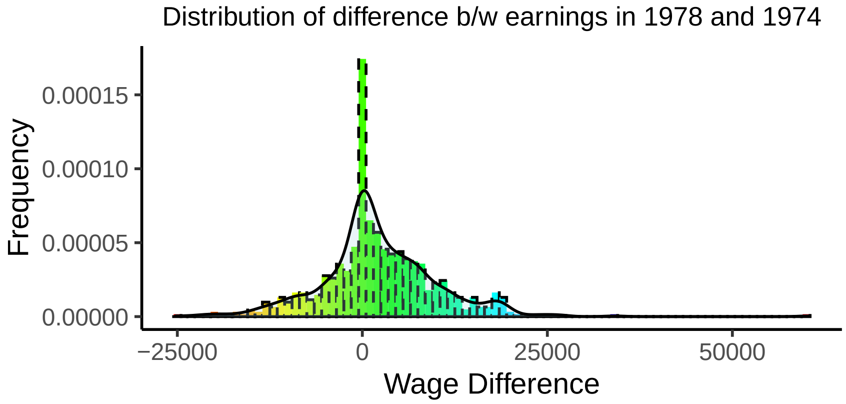
The data used in this analysis contains 614 male participants. The treatment group consists of participants for whom 1974 earnings can be obtained and the control group consists of all the unemployed males in 1976 whose income in 1975 was below the poverty level.

The real annual earnings for 1975 are measured during the course of the study and some participants were even paid during the course of the study. In order to ensure a better fit of our model and a sound analysis we chose to not incorporate the 1975 data in our analysis.

Since our question of interest is to evaluate whether participants who received training tend to earn higher wages, we created a new variable - $Wage\_difference$ based on the difference between real annual earnings in 1978 and 1974.

EDA:

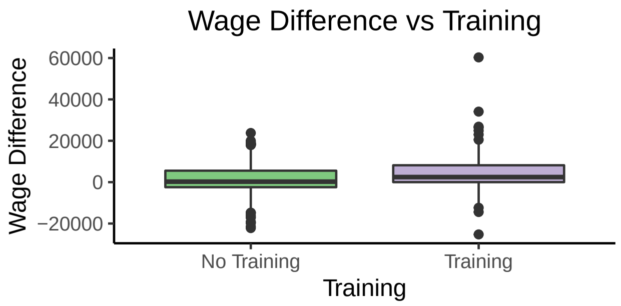
To proceed with our analysis, we plotted our response variable $Wage\_difference$ to check whether it follows a normal distribution in order to proceed with linear regression. We see that the distribution of wage difference is relatively normally distributed.

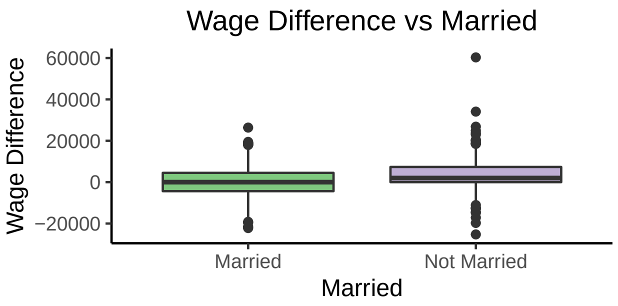


To explore the data further, we plotted our variables to establish any interesting associations between them as well as the response variable. Based on the scatter plots for the continuous variables we looked at whether there was any trend evident highlighting a specific relationship between these variables. The EDA was our first step in deciding which variables we should include in our model.

Based on a large number of plots and summary statistics, the indicator variables for whether or not the person was black, whether or not the person was hispanic, whether the person had a degree were identified as poor predictors. The boxplots for these variables did not indicate much of a difference based on category.

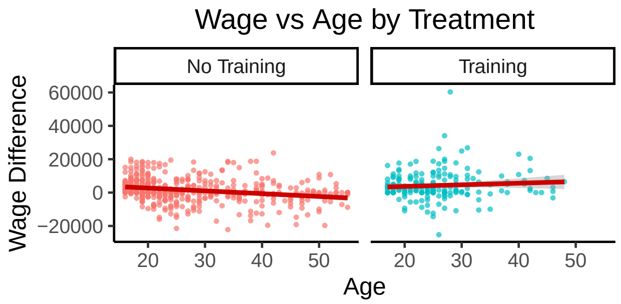
We did observe associations between wage difference and treatment and wage difference and married variables. There was a slight difference between the median values for people who received and did not receive training across these variables as can be seen in the plots below. We concluded these were important relationships to explore further.

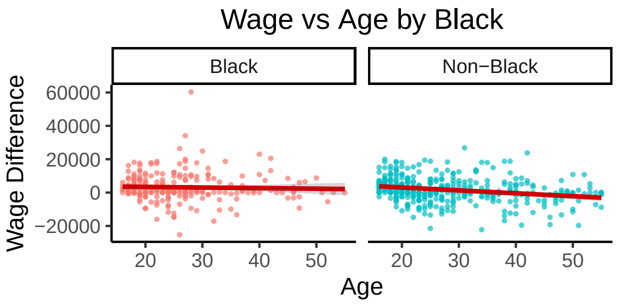




After assessing the relationships between the variables, we went on to further explore the interactions between variables. In particular, the boxplots for interaction between wage vs age by treat and wage vs age by Black seemed significant. Based on the scatter plots, we noted a difference in the trend of the scatter plot across training and concluded to explore these interactions further.

EDA Interactions





Model

To build our model, we first built our baseline model which included all our main effects. We then included some interaction effects which we thought were significant, or they answered questions with respect to the study. With the help of anova tests we assessed if they were significant to our model. Next we used to stepwise selection using AIC to to generate our final model.

Model Building

Our first model included the main effect of every variable. To improve our interpretation we also mean centered the continuous predictor Age. Here our response variable is **Wage Difference between 1974 and 1978** and the predictors are **treat, age (centered), black, hispan, education, married and no degree**. After building our model and assessing it, we quickly realized there was an outlier in our data set; observation 132. This point was not meeting our assumptions and it was showing up as a leverage point (more than 0.05) in our plot of Cook's Distance. The outcome of our model remains the same with an R-squared of 0.07 with or without this observation. Therefore, we decided to remove this observation from our model and go through the assessment again to ensure we did not necessarily violate any assumptions. Based on the summary of this model, we noted 3 significant variables: treat, centered age and married.

Next, we constructed another model using stepwise selection. A null model and a full model were determined, where the null model included variables of our interest (treat, black, hispan and age) and interactions between the demographics since they are questions of interest. The full model consisted of all variables the interactions between other demographics variables as well as education variable. We ran AIC as well as BIC to generate final models. The only difference between the two models was the variable married. We compared the 2 models using the anova test and choose AIC as our final model.

To ensure our final model is the best fit for our data, we also included the interaction effects we found interesting during our EDA to the model step by step and used the anova test to conclude whether they improved our model. There seemed to be no additional impact of those interactions and we moved forward with our final model generated by AIC.

Based on AIC our model our final model was:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} : x_{i2} + \epsilon_i ; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n.$$

where y_i is estimated value of difference in wages, x_1 is age centered, x_2 is treatment, x_3 is married, x_1:x_2 stands for the interaction of age centered and treatment.

Model Assessment

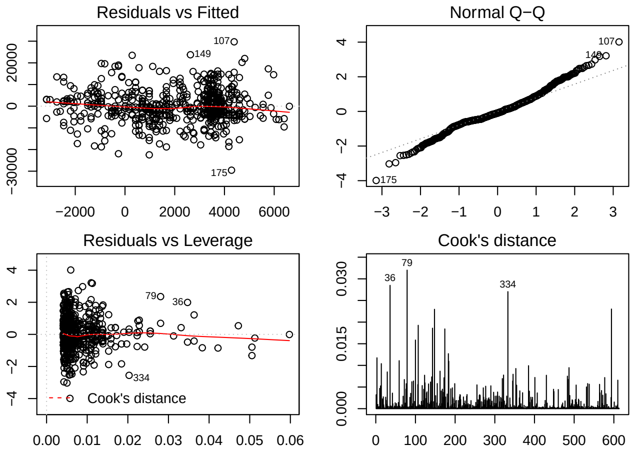
To assess our final model we checked if any of the assumptions of Linearity, Normality, Equal variance and Independence were violated.

In order to check the linearity assumption we plotted the residuals of the model against age. The points were randomly distributed and there was no visible pattern, therefore we can concluded that the linearity assumption is not violated.

To check the independence and equal variance assumptions we plotted the residuals against the fitted values. The points seemed randomly distributed with no discernible pattern and the spread of variables seemed constant above and below the line. There did seem to be some points on the x axis that may have violated the equal variance of errors assumptions however, they were only few and we went ahead and said that neither of the assumptions are violated.

To check for normality, we plotted the Q-Q plot. For our model we observed that majority of the points lie on the 45 degree line. There could be outliers present in the data, however there were not too many to say that the normality assumption was violated

We also looked at the leverage points and outlier and verified whether any of then were significant using cook's distance. All the points on the cook's distance plot were well below the 0.05 point and we concluded that there were no influential points.



We also checked if there was any multicollinearity between our variables to ensure that it did not hamper our analyses. The VIF values for all our variables were below 5 and we concluded there was no multocollinearity.

Model Interpretation

Here is the summary of our final model

Table 1: Linear Regression Summary

	Dependent variable:
	wage_diff
agec	−137.68*** (−207.34, −68.02)
treat1	2,123.82*** (762.35, 3,485.28)
married1	−1,652.88** (−2,996.46, −309.30)
agec:treat1	243.46*** (79.74, 407.17)
Constant	2,304.41*** (1,330.60, 3,278.22)
Observations	613
R^2	0.07
Adjusted R^2	0.07
Residual Std. Error	7,426.70 (df = 608)
F Statistic	12.00*** (df = 4; 608)
Note:	*p<0.1; **p<0.05; ***p<0.01

Based on our model, all our variables are statistically significant at the 95% confidence level.

Our response variable is Wage Difference and our explanatory variables can be interpreted as follows:

- β_0 is the intercept term. The intercept term gives us the average value of our response variable when the explanatory variables are 0. As per our model the the intercept value is 2304.41 which means that for a person of average age (27 years) who has received treatment and is married the wage difference is \$2,304.41.
- β_1 is the coefficient for the centered age variable. For a one year increase in age the wage difference reduces by \$137.68
- β_2 is the coefficient for the treat variable. For a participant who has received training, the wage difference between 1974 and 1978 is \$2,123.82 higher than someone who has not received training.

- β_3 is the coefficient for the married variable. For a participant who is married, the wage difference between 1974 and 1978 is $1,652.88 lower than a participant who is not married.
- β_4 is the coefficient for the interaction between centered age and treat variable. For a person who received treatment, one year increase in age will lead to an increase in the wage difference of $106.46.

The adjusted R^2 of our model is 7% which means that 7% of the variation in our model can be explained by our model. The standard error is 7426.70 with 608 degrees of freedom.

The 95% confidence interval for our variables is also included in our summary table. Specifically for the treat variable, our confidence interval is [762.35, 3485.28].

Conclusions

1. Based on our model, we can observe that participants who received training, had a higher difference in their annual earnings between 1974 and 1978. This could be concluded as evidence that participants who received training tend to earn higher wages, however we must consider the additional factors that may have influenced this. Our model also captures the interaction between age and treatment indicating that age may have a significant role to play in the impact training has on participant's wages.
2. Training is a statistically significant factor for difference in real annual earnings between 1974 and 1978. At the confidence level of 95% we see treat is significant and the coefficient can be interpreted as - For a participant who has received training, the wage difference between 1974 and 1978 is $2,123.82 higher than someone who has not received training. The confidence interval for the treat variable is [762.35, 3485.28] which implies that the difference between the real annual earnings for someone who received training against someone who did not lies in the range of $762.35 and $3,485.28
3. The demographic indicators pertaining to race (black, hispanic) were not significant and therefore were not included in our model. The demographic factor of age was found to be significant. Age has a negative impact on wage difference
4. We noted that there was an interesting association between married and wage difference during our EDA and it was also a significant factor in our model.

Potential Limitations

1. The linear regression model has very low r-squared value which limits the predictive accuracy of the model.
2. Our response variable is biased with a high presence of 0 values which may skew our results.

Designations: Aarushi Verma (*Writer*), Deekshita Saikia (*Checker*), Mohammad Anas (*Programmer*), Tego Chang (*Coordinator*), Sydney Donati-Leach (*Presenter*)

Part II

Summary

The focus of Part II is mainly to compare males who participated in the training program versus the ones who did not participate and determine whether the participants are more likely to be employed in 1978. We also explore demographic factors that are likely to increase the chances of earning non-zero wages for these workers. Exploratory data analysis is carried out on the dataset, and a logistic regression model is fit using stepwise selection. We observe that people who participated in the program are more likely to earn non-zero wages as compared to people who did not participate in the training program. The age of the male workers was a predictor of interest as its effect on the odds of earning non-zero wages was different for the males who participated as compared to the males who did not participate in training.

Data

The dataset consists of observations for 614 male workers, with 11 variables. We used the $re78$ variable available in the dataset to create a binary variable $re78Bi\_F$, which indicates whether the person was earning non-zero wages in 1978. We use this factor variable as the response variable in our analysis. We ensure that the independent variables have the correct data type before proceeding with the analysis. The variables *treat*, *black*, *hispan*, *married* and *nodegree* were converted to factors, while *age* and *educ* were used as numeric variables in our model.

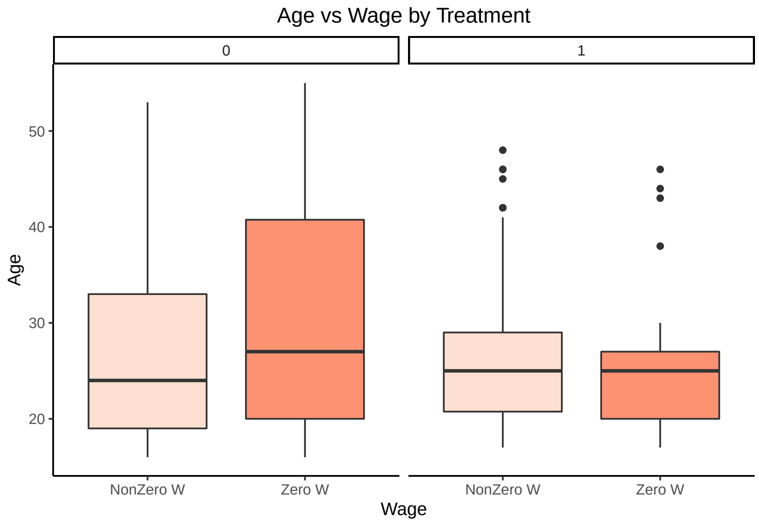
Exploratory Data Analysis

We start our exploratory analysis by observing the relationships of our predictor variables against our response variable, $re78\_F$. Given that the *treat* predictor is the main variable of interest, we explore it first. We see that the probability for earning non-zero wages is approximately the same for everyone regardless of whether they participated in the training program, as can be observed in the table below.

Table 2: Conditional Probabilities

	0	1
NonZero W	0.77	0.76
Zero W	0.23	0.24

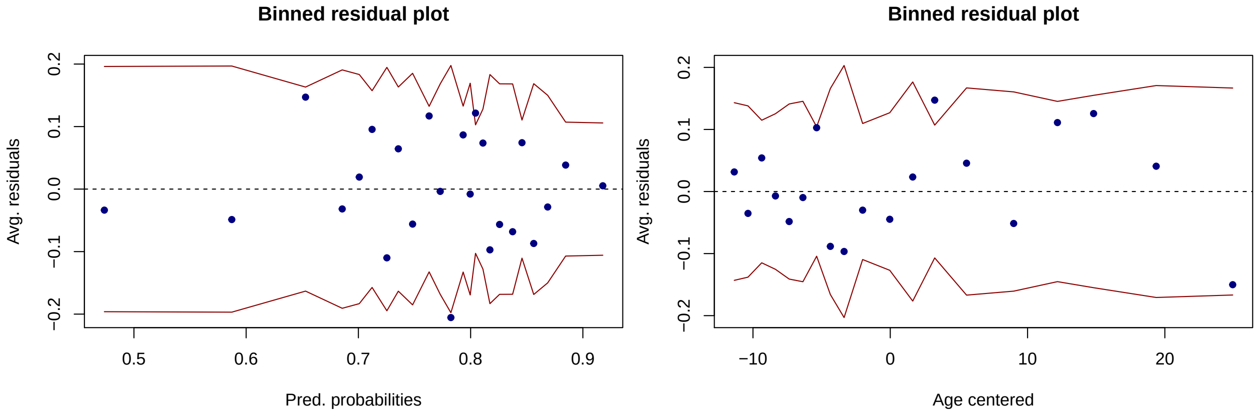
We also test the association of our response variable with the categorical predictors using Chi-squared test of association. We notice that black people are less likely to earn non-zero wages as compared to non-black people. The Chi-squared test shows us that the predictors *hispan*, *married* and *nodegree* have no association with the our response variable. We also generate box plots to explore the effects of our continuous predictors on our response variable. The box plots indicate that educated people are more likely to earn non-zero wages. We also note that younger people are more likely to earn non-zero wages. We move on to explore whether the effect of the predictors on our response variable may be affected by other predictors. The box plot below indicate that for people who did not participate in the NSW training, younger people are more likely to earn non-zero wages, while the trend seems to be the opposite for the people who participated in the training program.



We also observed that for non-black people, the training was more effective. Therefore, we include the interaction for the *black* and *treat* predictors in our model. Based on the results from our exploratory analysis, we saw that the interaction affects of *age* and *educ*, *hispan* and *treat*, and *re*74 and *treat* might be worth investigating as well.

Model

We start by fitting a model that includes only the main effects except the *nodegree* variable. We exclude this variable from our model as it captures very similar information to the *educ* variable. The results of the model seem counter-intuitive as our main variable of interest, *treat* is statistically insignificant. We also note that the real earnings of a person in 1974, *re*74, the centered age variable, *age_c* and *black* predictors have a significant affect on the odds on earning a non-zero wage in 1978. The residual deviance of our model is 634.95 which suggests that model is a better fit than the null model. To assess the model further, we observe binned plots of the residuals against the fitted values and the continuous predictors. We check for randomness in these plots to ensure that our model satisfies the independence of errors assumption and to investigate whether any transformations of the continuous predictors are required. The binned plot for residuals against fitted values seem fairly random, except for a couple of bins on the left area of the plot. The binned plots against continuous predictors look random for *educ* and *re*74, indicating that we do not need any transformations for these two variables. However, for *age_c*, we see a polynomial trend that has not been captured by our model.



To improve our model's fit, we start adding interactions between the effects to our model. We adopt a stepwise variable selection approach combined with Chi-squared tests to see which predictors and interaction effects amongst the predictors improve the fit of our model. To do this we specify a base/null model, where we only include the variables of interest; the treatment variable *treat*, demographic variables (*black*, *hispan* and *age_c*) and the interaction of *treat* with all the demographic variables. On the other hand, our full model contains all the variables and interactions in the null model, the interactions that were found to be interesting in our exploratory analysis, and the *married* and *educ* effects. We then perform stepwise model selection, using AIC as well as BIC as the criteria for variable selection. Both iterations of the logistic regression models shortlist *treat*, *age_c*, *re*74 and the interaction between *treat* and *age_c* variables in the stepwise selection process. However, as AIC tends to be more lenient as compared to BIC, we observe that *black*, *hispan*, interaction between *treat* and *hispan*, and the interaction between *treat* and *black* variables are also selected in the model using AIC for variable selection. The results of the model containing predictor variables using AIC are shown below.

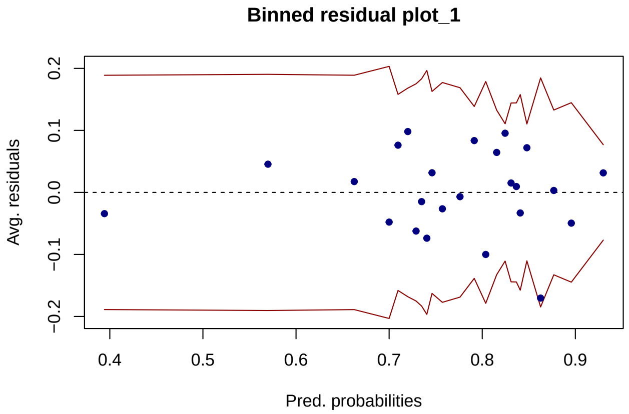
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.98	0.18	5.31	0.00
treat1	0.52	0.29	1.78	0.08
age_c	-0.06	0.01	-4.95	0.00
black1	-0.48	0.27	-1.79	0.07
hispan1	0.05	0.36	0.13	0.90
re74	0.00	0.00	3.89	0.00
treat1:hispan1	15.09	722.55	0.02	0.98
treat1:age_c	0.08	0.03	2.86	0.00
treat1:re74	-0.00	0.00	-1.68	0.09

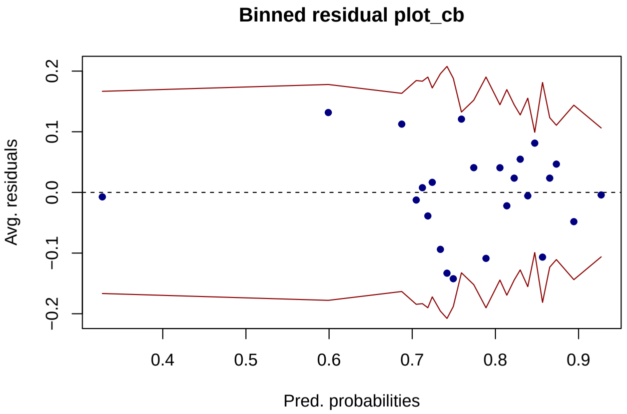
Table 3: Logistic Regression Results (Log Odds Scale)

We observe that not all the predictors in the AIC model are significant at the 5% significant level. To check which iteration of the model fits better, we conduct a Chi-squared test and compared the residual deviance of these two model iterations. The p-value of this test turns out to be 0.012, indicating that at least one of the additional variables selected in our AIC model is significantly improving the fit of the model. We see in the results of our AIC model that the variables *black* and the interaction between *treat* and *re*74 have low p-values. Therefore, we fit a new model which included these two effects and all the variables from our BIC model. We compare this model to the BIC model using a Chi-squared test. The difference was statistically significant and we decided to keep these effects in our model. To check if the remaining effects in the AIC model (*hispan*, and its interaction with *treat* variable) would improve the fit of our model, we conduct another Chi-squared to compare the residual deviance between our new model and the AIC model. Based on the results of this test, which yields an insignificant p-value, we conclude that these effects do not lead to a statistically better fit.

The results of our model suggest that all the variables are significant at the 5% significance level except the interaction of *treat* and *re*74 variable. We also note that the residual deviance of this model falls to 626.03 proving that this is a better fit as compared to our baseline model that included only the main effects. However, when we look at the binned plot of residuals against our continuous predictors and predicted probabilities, we see that the polynomial trend in the *age_c* variable has still not been captured.

We observe that there is still a trend in the residuals plot of the *age_c* variable, which prompts us to investigate polynomial interactions of this effect in the model. We explore squared and cubic degree polynomials of *age_c* in our model. We also included the interactions of the *treat* variable with these transformations of *age_c*. We run the regression again with the additional variables to check if this improves our model fit. However, the residual deviance of this model turned out to be 620.32 which is not a significant improvement over our previous model, which is also confirmed by a Chi-squared test. We compared the binned residual plots of these models against the fitted values. These plots are shown below.



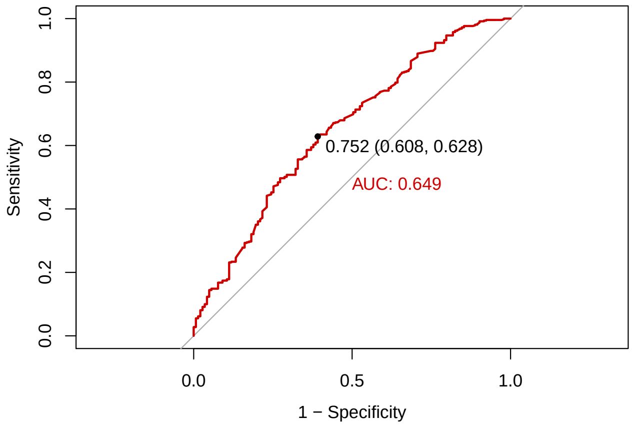


We see that there is not much difference in these plots and the bins on the far left of the plot are still present. Given that adding the squared and cubic terms of *age_c* in our model makes it harder to interpret the standard estimates and does not significantly improve the fit of our model, we decided not to include them in our model. Now that we have investigated the effects of including interactions and transformations to our model, our final model equation is given below.

$$y_i|x_i \; Bernoulli(\pi_i); log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i1} : x_{i2} + \beta_6 x_{i1} : x_{i4}$$

Our y value is the binary variable to indicate whether a participant earned a positive wage. Our x values are the treatment, age centered, black (factor), re74, interaction of treatment and age centered and interaction of treatment and re74.

We now assess model performance by observing the RoC curve, as shown below.



We then leverage the confusion matrix to calculate the accuracy, sensitivity and specificity of our model. Classifying the outcomes using the probability threshold of 0.5 allows us to achieve an accuracy of 78% and sensitivity of 98%. However, we see that our model does a poor job at predicting people who earned zero wages in 1978 (as can be seen from the low specificity of 11%). Therefore, to obtain a balance between specificity and sensitivity, we classify outcomes based on the probability threshold suggested by the ROC curve, which is 0.752. Using this threshold, we are able to achieve sensitivity rate of 63% and the specificity rate improves to 60%.

The standard estimates of all the coefficients in our model were statistically significant except the interaction between the *treat* and *re*74 predictors. We exponentiate the standard estimates and their confidence intervals to interpret them on the odds scale. The model results are as shown below.

Despite having interactions, multicollinearity is not present in our model as the variance inflation factor of each of the variables is well within range (<10).

	coeffecients_2.5	coeffecients	coeffecients_97.5	Std. Error	z value	Pr(>|z|)
(Intercept)	1.99	2.77	3.87	0.17	5.99	0.00
age_c	0.92	0.95	0.97	0.01	-4.98	0.00
re74	1.00	1.00	1.00	0.00	3.84	0.00
treat1	1.08	1.89	3.31	0.29	2.24	0.03
black1	0.34	0.55	0.90	0.25	-2.40	0.02
age_c:treat1	1.02	1.08	1.14	0.03	2.77	0.01
re74:treat1	1.00	1.00	1.00	0.00	-1.60	0.11

Table 4: Logistic Regression Results (Odds Scale)

Conclusion

To sum up our analysis, the people who participated in the program are more likely to earn non-zero wages in 1978 as compared to people who did not participate. Statistically speaking, the odds of earning positive wages after receiving training is 1.89 times the odds of a person who did not receive training for a 27 year old male who was unemployed in 1974. A likely range of the odds for the effect of training was found to be $[1.08, 3.31]$. Age was found to an interesting variable in our analysis. For a person who did not receive training, a 1 year increase in age decreases the odds of earning positive wages by 5.5%. The association between *age_c* and *treat* was also an interesting association in the model. For people who participated in the training, 1 year increase in age leads to an increase in the odds of earning non-zero wages by 1.9%. We get this value by multiplying the standard estimates (odds scale) of *age_c* and the estimates of the interaction between *age_c* and *treat*. Non-black people are also more likely to earn non-zero wages in 1978.

Limitations

1. Our model assumes that the length of the NSW training remained the same for all participants. However, in reality, participants joined the training program between March 1975 and July 1977, and the randomization over this 2-year period led to people with different characteristics joining the program.
2. There is an uneven distribution in our response variable. The proportion of people earning non-zero wages is very high compared to the proportion of unemployed people. A potential solution to this problem can be under sampling the data of males earning non-zero wages.

Designations: Aarushi Verma (*Coordinator*) ,Deekshita Saikia (*Programmer*), Mohammad Anas (*Writer*), Tego Chang (*Checker*), Sydney Donati-Leach (*Presenter*)