Data Analysis Team Project 1

Jeremy Zeng, Dean Huang, Yuwei Zhang, Zihao Lin

9/27/2020

Part I

Summary

The goal of this report is to address the question "Is there evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training?" In other words, we are trying to measure the effect of the treatment on real annual earnings by calculating the change of the wage between 1974 to 1978 for workers. Scatter plot will be used to analyze the association between the numeric response variable and the numeric predictor variable, and boxplot will be used to analyze the associations between numeric response variables and the categorical predictor variables. Residual plots and Q-Q plots will be used to verify the assumptions of the final regression model. VIF will be used to calculate the multicollinearity of the model. The outcome of the study shows that workers who receive job training tend to earn more than workers who do not receive job training, and there is enough evidence to conclude that the association between treatment and wages differs by the worker's age.

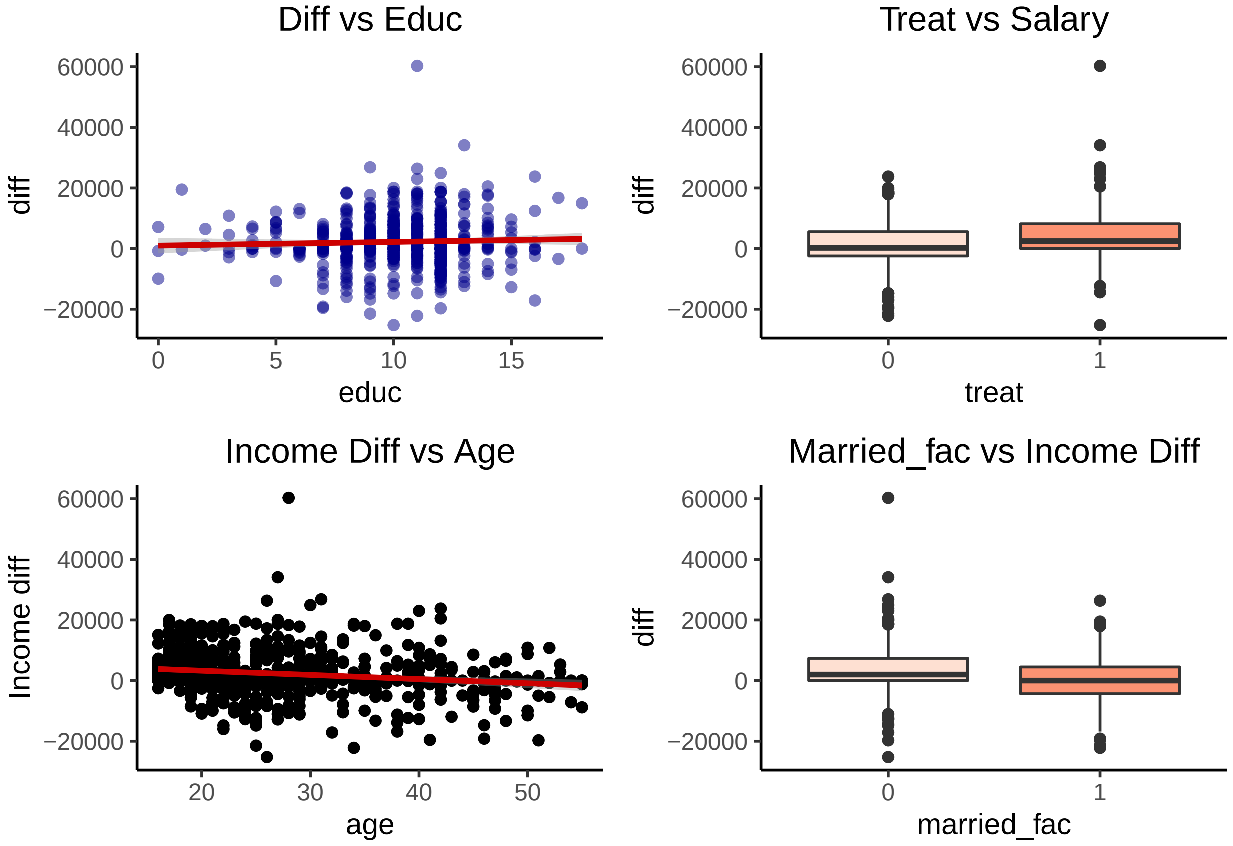
Introduction

The response variable we are interested in studying is the change of workers' wages from 1974 to 1978, and the predictor variables we are interested in looking to are treat, age, degree, years of education, marriage status, black, and hispanic. After completing data modeling and model validation, the estimated range for the effect of training on wages will be calculated from the final model. Interesting associations with wages will also be highlighted. Our study will begin with EDA with the goal of checking the association of predictor variables and the response variable, and highlight the preliminary concerns we have based on the results of EDA. Through careful analysis of the results from EDA, transformation will be performed and interactions will be added to improve the fit of the model. Stepwise selection method will be performed to find the optimal model with the lowest AIC score. Before dropping the predictor variables and interactions, F-test will be implemented to assess the impact of variables/interactions on the predictive model. Last but not least, we will check if our final model fulfills the linear regression assumptions, and do not have multicollinearity and outliers.

Data

EDA of Response Variable & Predictor Variables

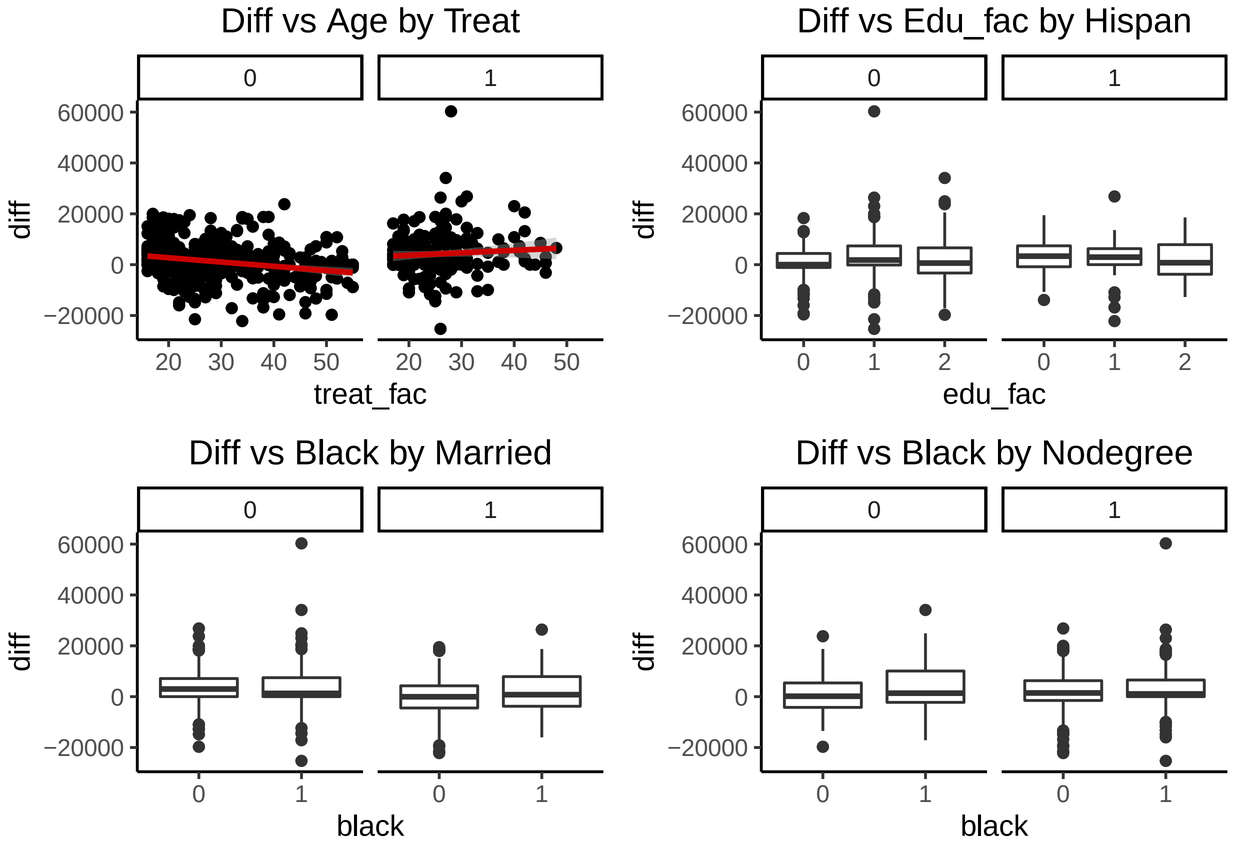
Both the histogram and Q-Q plot of wage difference appear to be normally distributed; however, there are some points deviated from the 45 degree line. Therefore, the transformation might be needed to accurately model the relationship between wage difference and predictor variables. Log transformation of the response variable will result in significant interpretation error because log cannot take in negative values. In addition, if we square the wage difference, we will not be able to differentiate if the difference is positive or negative. The scatter plot of wage difference vs education shows that most points are between 5-15 years, and there is insufficient data for education fewer than 5 years and education more than 15 years. Therefore, we have decided to categorize education into different levels according to the American education system and the sheepskin effect (people possessing an academic degree earn a greater income than people who have an equivalent amount of studying without possessing an academic degree). The three education levels are: Below High School (years of education lower than or equal to 9), High School Incomplete (years of education between 9 and 11), High School & Above (years of education higher than or equal to 12). The boxplot of wage difference vs education levels shows there is slight difference in distribution and median for different education levels. According to the boxplot and box plot of treat vs wage difference, those who received job training appear to have slightly higher positive wage differences (more investigation is needed). The boxplot of black vs wage difference shows the wage difference distribution for black is slightly higher than the wage distribution of others. The boxplot of marriage status vs wage difference shows married individuals appear to have lower wage differences than singles. The boxplot of hisp vs wage difference shows the wage difference distribution for hispanic is slightly higher than the wage distribution of others. The boxplot of nodegree vs wage difference shows similar distribution in all groups but the median for group 1, who dropped out of high school, is higher than other groups. The scatter plot shows some association between age and wage difference; however, there is insufficient data for people above 40 years old.



Interaction

Treat Interactions: Since we are interested in assessing the association between treat and wage difference, we will investigate the interactions of treat with other predictor variables. Scatter plot of wage difference vs age by treat shows difference in trend between the group that receive job training and the group that receive no job training (negative trend with no treat and positive trend with treat). The box plot of wage difference vs edu by treat shows difference in trends of treat 0 and treat 1 for different education levels. The box plot of wage difference vs marriage status by treat shows married individuals tend to have higher positive wage differences than single individuals with training; the outcome is opposite for the without training group. The box plot of wage difference vs black by treat shows individuals who are black and have job training appear to have lower positive wage difference than individuals who are not black and with job training; however, the trend is opposite for non-training groups. As for hisp and nodegree, there appear to be no obvious differences in trends for different treat groups.

Other Interactions: According to the scatter plots of wage difference vs age by black, education level, hisp and married, there appear to be no interactions between age and the categorical variables listed above. The box plot of wage difference vs age by treat shows a positive trend after training compared to a negative trend without training, indicating interaction. For education level, the box plot of wage difference vs education level by black, married and nodegree show no difference in trend, indicating no interaction. However, the box plot of wage difference vs education level by hisp has a slightly negative trend in hispanic group and a bell-shaped curve in non hispanic groups, indicating some interaction. For categorical predictor black, the box plot of wage difference vs black by married shows a positive trend in married group and a negative trend in non-married group. In addition, the box plot of wage difference vs black by nodegree shows a negative trend in nodegree group compared to positive trend in degree group, indicating some interaction. For categorical predictor hisp, the box plots of wage difference vs hisp by married and nodegree show no difference in trend. For categorical predictor marriage, the box plot of wage difference vs marriage by nodegree shows no difference in trend.



Model

Stepwise method will be implemented to find the lowest AIC because BIC generally places a heavier penalty on modes with more than 8 variables. The single predictor variables we will include for our full models are treat, age, educ, married, black, hisp. and nodegree. Besides the single predictor variables, we will include ten interactions including 1) treat and age 2) treat and educ 3) treat and married 4) treat and hisp 5) treat and black 6) treat and nodegree 7) educ and hisp 8) black and nodegree 9) black and married. After performing stepwise selection, the final model ended up having four predictor variables: married, treat, age, and treat*age. Through the results of EDA and potential interaction investigation, six other potential interactions are identified: 1) educ vs income difference by treat 2) marriage vs income difference by treat 3) nodegree vs income difference by treat 4) education vs income difference by hispanic 5) black vs income difference by marriage 6) black vs income difference by nodegree. F-test is conducted to decide whether to drop the interactions, and the result of f test shows including these interactions have high p-values compared to excluding these interactions. The next step is to check the multicollinearity of the final model, and all vif values are below 5, which is acceptable. For the linearity assumption, the residual plot for income difference vs age appears to be linear because the plot seems random. The residual fitted plot looks random and "roughly" equally spread out around zero. Therefore, no violation of the independence and equal variance assumption. Most points appear to cluster around the 45 degree line of the Q-Q plot with some points at both ends of tails deviating from the 45 degree line. Overall, the model satisfies the normality assumption. In addition, there are no influential points and outliers according to the graph of Cook's distance. However, there are some leverage points that are not influential.

$$y_i = \beta x_i + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

This is the equation of our final model. y_i is the wage difference in dollars for observation i, and x_i is the vector containing the corresponding values for marriage, age, treat, and age:treat.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2356.2993	518.8001	4.54	0.0000
married_fac1	-1756.5158	715.7173	-2.45	0.0144
age_cen	-135.7926	37.1147	-3.66	0.0003
treat_fac1	2415.3552	724.2527	3.33	0.0009
age_cen:treat_fac1	255.8791	87.2117	2.93	0.0035

	2.5 %	97.5 %
(Intercept)	1337.44	3375.15
married_fac1	-3162.09	-350.94
age_cen	-208.68	-62.90
treat_fac1	993.02	3837.69
age_cen:treat_fac1	84.61	427.15

Conclusion

According to the summary table of our final model, predictors that are significant in the 95% confidence interval are married, age, treat, and age:treat, with p-values lower than 0.05.The intercept of the final model shows an individual who has not received training, with an average age of 27, and who is single will have a wage difference of 2356.30. As the marriage status changes from single to married, the wage difference will decrease by 1756.52 with all other variables constant. As the age increases by one, the wage difference will decrease by 135.79 with all other variables constant. As treat changes from 0 to 1, the wage difference will increase by 2415.36. With 95% confidence, the range of the increase of wage difference of receiving treat is between 993.02 and 3837.69 with other variables constant. One more year in age for individuals who receive job training will lead to a 255.88 additional increase in wage, holding all other variables constant. The adjusted R square of 0.07 means 7% of variation in the response variable is explained by the regression fit. One potential limitation of the model is that wage does not necessarily reflect the total income you earn because there is a possibility that some incomes are not reported (Eg. e-commerce).

Part II

Summary

The goal of this report is to address the question "Is there evidence that workers who receive job training tend to be more likely to have positive (non-zero) wages than workers who do not receive job training" In other words, we are trying to measure the effect of the treatment on helping individuals acquire a job. Boxplots and binned plots will be used to analyze the association between binary response variable and numeric predictor variable. Joint probability table and conditional probability table will be used to analyze the association between binary response variable and categorical predictor variable. Binned residual plot will be used to evaluate the overall fit of the regression model, and check if the function of numeric predictors is well specified. Chi-squared test will be used to compare the deviance of null model and new model. Confusion matrix and ROC curve will be used to validate the performance of the model through calculation of sensitivity, specificity, accuracy, and AUC curve. VIF will be used to calculate the multicollinearity of the function.The outcome of the study shows there is not enough evidence to conclude that workers who receive job training have higher odds of having non-zero wages. However, black, education, age^2, and age:treat are important predictors for the odds of having non-zero wages because the p-values are lower than 0.1.

Introduction

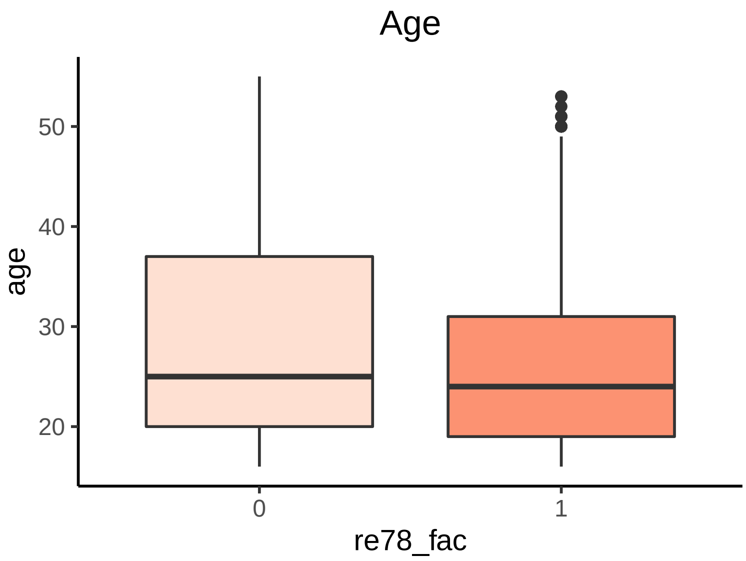
The response variable we are interested in studying is the workers' wages in 1978, and the predictor variables we are interested in looking to are treat, age, degree, years of education, marriage status, black, and hispanic. After completing data modeling and model validation, the estimated range for the difference in odds ratio of having non-zero wages for treat 0 and treat 1 will be calculated from the final model. Interesting associations with odds of non-zero wages will also be highlighted. The study will begin with EDA with the goal of checking the association of predictor variables and response variables, and highlight the preliminary concerns for the response and predictor variables. Next, preliminary logistic model fitting will be performed (excluding interactions and transformation) to understand the significance of coefficients for each selected predictor variable. Binned residual plots will be utilized to assess the overall fit of regression model and check if the function of predictors is well specified. Next, preliminary model validation will be performed to understand the fit of the model for making predictions. Confusion matrix and ROC curve will be utilized to validate the performance of the model through calculation of sensitivity, specificity, and accuracy. Through careful analysis of the results from EDA, preliminary model fitting and preliminary model validation, transformation will be performed, and interactions will be added to improve the fit of the model. Logistic model fitting and model validation processes will be performed again to justify each modification made to the model. In addition, chi-squared test will be implemented to compare the performance of the new model to the original. Lastly, stepwise function will be performed to find the optimal model with the lowest AIC score. To ensure the final model fulfills the assumptions of logistic regression, model validation will be performed again. Last but not least, VIF will be used to check for any multicollinearity for the final model.

EDA

The first step is to categorize the response variable, re78. If the value of re78 is larger than zero, we will set it to 1, otherwise, we will set it to 0. In our dataset, the total number of data that is 0 is 143, and the total number of data that is 1 is 471. There are 7 predictors: treat, education, black, hisp, married, nodegree and age. The first six variables are categorical variables. The scatter plot of wage difference vs education shows that most points are between 5-15 years, and there is insufficient data for education fewer than 5 years and education more than 15 years. Therefore, we have decided to categorize education into different levels according to the American education system and the sheepskin effect (people possessing an academic degree earn a greater income than people who have an equivalent amount of studying without possessing an academic degree). The three education levels are: Below High School (years of education lower than or equal to 9), High School Incomplete (years of education between 9 and 11), High School & Above (years of education higher than or equal to 12). We will begin our EDA by analyzing the association of the response variable and categorical predictors: treat, education, black, hispan, married and nodegree. Three tables are drawn for each predictor. The first table shows the number of positive salaries and the number of zero salaries for each level of categorical predictor. The second table shows the probability of each combination of response variable and categorical predictor with the denominator as the total number of data. The third table shows the conditional probability for each combination of response variable and categorical predictor. Next, we calculate the p-value of Chisq-test to see if the predictor is significant. According to the p-value table below, only black and education variables seem to have association with response variable (more investigation is needed).

variable	treat	education	Married	black	hispan	nodegree
p-value	0.7686	**0.0113**	0.5756	**0.0201**	0.2052	0.5053

Next, we will analyze the association of the response variable and numeric variable, age. According to the box plot and binned plot, there appears to be some association between age and salary.



To sum up, education, age and black appear to have some association with salary.

Model

After conducting EDA, the next step is to construct the preliminary logistic model and conduct model validation with all major predictor variables for this study (excluding transformation and interaction), which are treat, age, degree, years of education, marriage status, black, and hispanic. The summary table of the preliminary model shows that only the coefficients of race (black) and age are significant, the rest of the predictor variables have a p-value above 0.05. In addition, according to the binned residuals versus predicted probabilities plot, all points except two are within the standard error bounds and the overall plot appears to be random. For age, the binned residual plot displays a quadratic pattern with all points except two within the standard error bounds (transformation appears to be needed). According to the confusion matrix of the preliminary model with 0.5 threshold, the optimal sensitivity and specificity is (0.992, 0.063) and the overall accuracy is 0.78. In addition, the ROC curve shows the optimal 1-specificity and sensitivity is (0.82, 0.36) and the AUC value is 0.628, which is not ideal. Improvement of the preliminary model is required.

Through analysis of results of binned residual plots, transformation for age appears to be necessary because the plot exhibits a quadratic trend. Both log transformation and quadratic transformation are performed on age, but only the binned plot of quadratic transformation produces better results. After performing quadratic transformation, the shape of the binned plot for age appears to be more random. We will perform model fitting again but this time adding age^2 to the model. According to the binned residuals versus predicted probabilities plot, all points except one are within the standard error bounds, and the overall plot appears to be random. Outliers still exist but it is better than the previous binned residual plot. The binned residual plot for quadratic age appears to be more random. The confusion matrix with 0.5 threshold shows the accuracy is 0.78 and the optimal sensitivity and specificity is (0.993, 0.077), which is a little bit better than the results of the preliminary model. The AUC value of the ROC curve is still 0.629, which is only 0.001 better than that of the preliminary model. Moreover, the age square variable has a p-value of 0.049, which is lower than 0.05. Since the p-value of the chi-squared test is 0.049, we will include the age square variable to our model.

The next step will be investigating potential interaction in logistic regression. Since the main purpose of this study is to find the association of job training and wages, interaction of treat and all other predictor variables will be investigated. Binned plot for numeric variable age vs wages by treat shows a difference in distribution between workers who receive job training and workers who do not receive job training. According to Pearson's Chi-squared test results, the p-value for education vs wages by non-treat group is 0.008 which is smaller than 0.05. The p-value of Pearson's Chi-squared test for education vs wages by non-black group is 0.006 which is smaller than 0.05. Therefore, the interaction of treat and education, and the interaction of black and education appear to be potential limitations.

For this report, stepwise selection method will be implemented to find the lowest AIC because BIC generally places a heavier penalty on models with more than 8 variables. Main predictors that will be included to the full models are treat, years of education, age, black, hisp, marriage status and degree. Besides the main predictors, we will include 10 interactions, including: 1) treat vs all other variables 2) education and black 3) black and marriage status 4) black and degree 5) age and education. After performing stepwise selection, the final model ended up with five predictors: age(centered), age^2, treat, educ, and black. These five predictors match the finding from EDA. However, all interactions are dropped from the model. Through the results of EDA and potential interaction investigations, three potential interactions and transformation of age are identified: 1) education vs wages by treat 2) education vs wages by black 3) age vs wages by treat 4) age vs wages by education. Chi-squared tests are conducted to decide whether to include the interactions and the square of age in the final model. F-test is conducted to decide whether to drop the interactions, and the result of f test shows including age:treat and age:education have a low p-value compare to excluding these interactions. Therefore, we have decided to add these two interactions into our final model. The binned residual plots of our final model look random, and all points are within the standard error bound. In addition, all VIF values are smaller than 10, which is good (no violation of multicollineaity). According to the confusion matrix with 0.5 threshold and ROC curve of the final model, the optimal sensitivity and specificity is (0.994, 0.077), accuracy is 0.78, and AUC value is 0.653, which are still not ideal but better than the preliminary model.

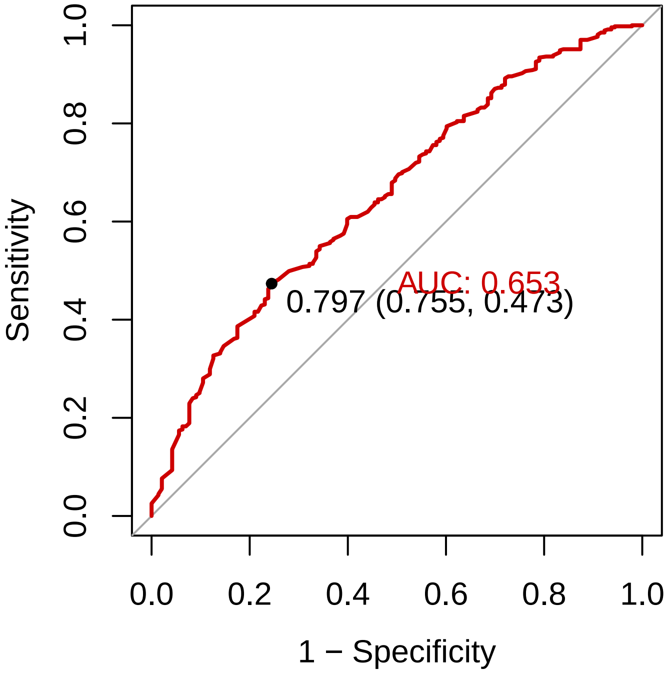
$$log(\frac{\pi_i}{1-\pi_i}) = \beta x_i; Bernoulli(\pi_i).$$

Above is the equation of our final model. pi/(1-pi) is the odds of positive wages for observation i, and x_i is the vector containing the corresponding values for treat, age(centered), age^2, education, and black.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.3690	0.2544	5.38	0.0000
age_cen	-0.0067	0.0210	-0.32	0.7490
black_fac1	-0.6913	0.2511	-2.75	0.0059
treat_fac1	0.1601	0.2774	0.58	0.5639
edu_fac1	0.5237	0.2669	1.96	0.0497
edu_fac2	0.3383	0.2650	1.28	0.2018
age_sq2	-0.0024	0.0010	-2.33	0.0195
age_cen:treat_fac1	0.0478	0.0275	1.74	0.0821
age_cen:edu_fac1	-0.0265	0.0243	-1.09	0.2743
age_cen:edu_fac2	0.0276	0.0231	1.20	0.2321

The table shows the summary of model including the interactions of age with treat and education.

	2.5 %	97.5 %
(Intercept)	0.88	1.88
age_cen	-0.05	0.03
black_fac1	-1.19	-0.20
treat_fac1	-0.38	0.71
edu_fac1	-0.00	1.05
edu_fac2	-0.19	0.86
age_sq2	-0.00	-0.00
age_cen:treat_fac1	-0.00	0.10
age_cen:edu_fac1	-0.07	0.02
age_cen:edu_fac2	-0.02	0.07



Conclusion

According to the summary table of our final model, predictors that are significant in the 90% confidence interval are black, education, ageˆ2, and age:treat, with p-values lower than 0.1. The intercept of the final model shows an individual who has not received training and is not black, with an average age and an education level below high school has an odds of having positive wages of 3.93%, exp (1.37). As treat changes from 0 to 1, the odds of having positive wages will increase by approximately 17.3% with all other variables constant. Since the p-value of treat is 0.56, the impact of treat on the odds of having positive wages is not significant. Compared to individuals who are not black, the odds of having positive wages will decrease by 49.9% (1-exp(-0.69)). With one unit increase in the square of centered age, the odds of having positive wages will decrease by 0.24%, holding all other variables constant. Compared to level 0 education (below high school), one level increase of education (high school incomplete) will lead to an increase of 68.8% in the odds of having positive wages. According to the interaction of age:treat, as the age increases by one, the odds of having positive wages will increase by an additional 4.8% (1- exp(0.05)) on top of the effects of age and treat predictors. With 95% confidence, the range of the increase of odds of having job training compared to individuals who do not receive job training is between -31.6% and 100.3% with other variables constant.

The range of increase of the odds for the black group are 69.4% and 18%. One potential limitation of the model is that wage does not necessarily reflect the total income you earn because there is a possibility that some incomes are not reported (Eg. e-commerce). In addition, the model does not take into consideration the retirement status of individuals; it is possible that individuals with wage 0 retired in 1978. Therefore, non-zero wage does not always reflect the employment status of the individuals.