Exploratory Data Analysis

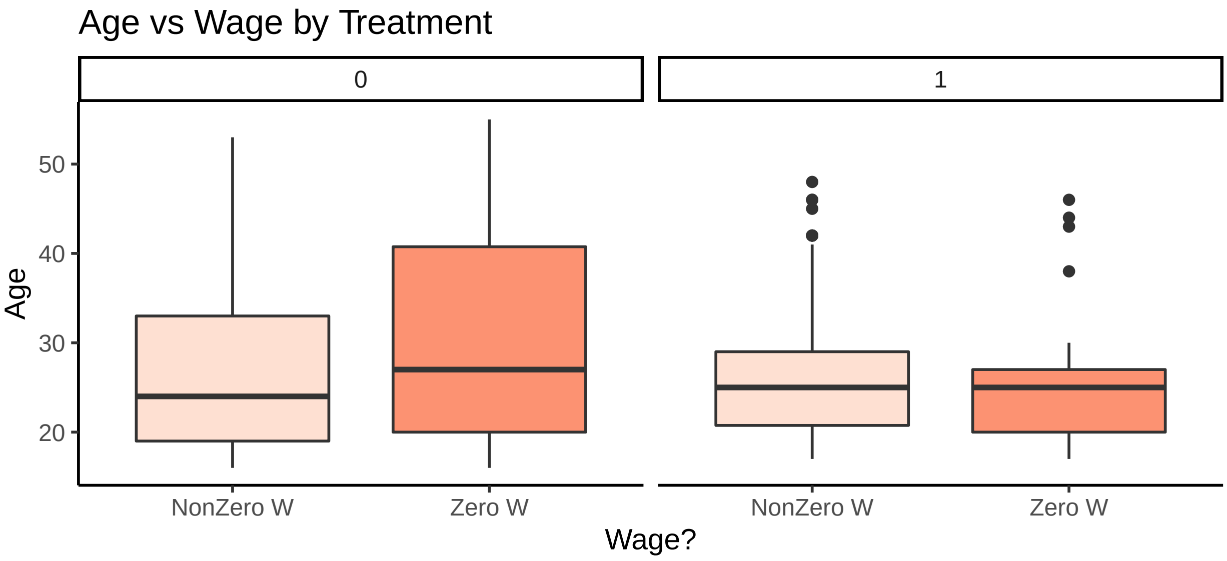
We start our EDA by observing the trend of our preditor variables on our response variable which is re78_F. Given that the treat variable is the main variable of interest, we explore it first. We see that the probability for earning a non zero wage is approximately the same for everyone regardless they took part in the NSW training or not. We also conducted Chi-squared test to test the association between the two variables and given our p-value, we fail to reject the null. The table indicating the conditional probabilities is provided below:

Table 1: Conditional Probabilities

	0	1
NonZero W	0.77	0.76
Zero W	0.23	0.24

We test the association of our response variable with other factor variables as well using the similar procedure. We notice that black people are less likely to earn a non-zero wage as compared to non-black people. The Chi-squared test shows us that the variables Hispanic, married and no-degree have no association with the our response variable. We create box plots to explore the affect 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.

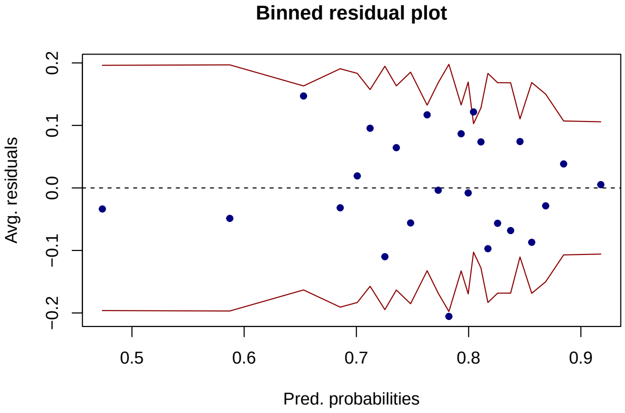
We move on to explore whether the effect of any of our predictors on our response variable varies with other variables. The box plots below indicate that for people who did not participate in the NSW training, younger people are more likely to earn non-zero wages, while the trend seems to be the opposite for the people who took part in the training.

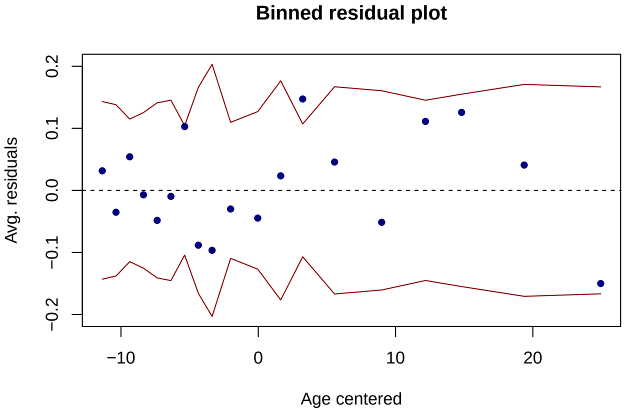


We also observed that for non black people, the training was more effective. Therefore, we include the interaction for black and treatment variable in our model. Based on the results from our EDA, we saw that the interactions affects of age and education, Hispanic and treatment variable and earning in 1974 and treatment variable are worth investigating as well.

Model Building

We start by fitting a simple model that includes only the main effects except the no-degree variable. We exclude this variable from our model as it is strongly correlated with our education variable. The results of the model seem counter-intuitive as our main variable of interest, the treatment variable is statistically insignificant. We also note that the real earnings of a person in 1974, age_centered and black variables 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. Moving on to the model assessment, we make 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 we need to add transformations for the continuous predictors. Our binned plot for residuals against fitted values looks random except that there are two outliers on the left side of plot. Our binned plots against continuous predictors look random for education and re74, indicating that we do not need any transformations for these two variables. However, for the age_centered variable, we see a polynomial trend that has not been captured by our model.





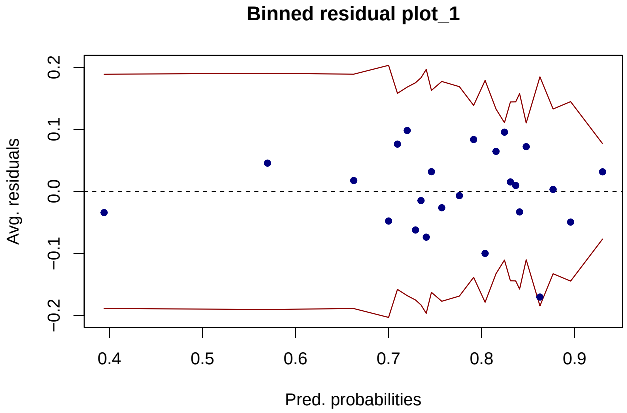
To improve our model's fit, we start adding interactions to our model. We adopt a step wise variable selection approach combined with Chi-squared tests to see which variables and interaction affects improve the fit of our model. To do this we create a null model, where we only include the variables of interest; the treat variable, demographic variables (black, hispanic and age_centered) and the interaction of treat variable 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 the EDA and the married and education variable. We then perform step wise model selection twice, once using AIC as a criteria for variable selection and once using BIC. Both AIC and BIC allow us to keep treat, age_centered, re74 and the interaction between treat and age_centered variables in our model. However, as AIC tends to be more lenient as compared to BIC, it also selects black, hispanic, interaction of treat and hispanic variable and interaction of treat and black variables as well. The results of the model containing predictor variables using AIC are shown below.

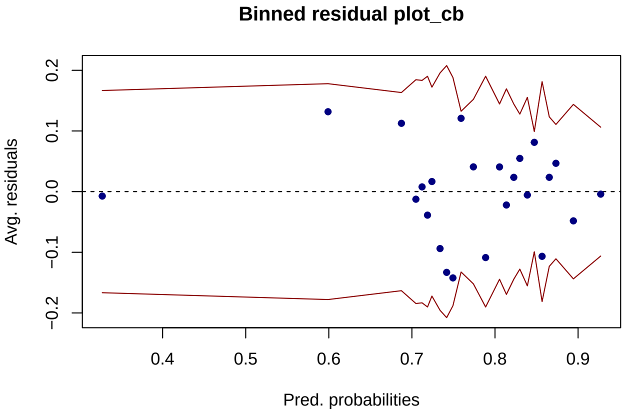
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.98	0.18	5.31	0.00
treat1	0.52	0.29	1.78	0.08
age_c	-0.06	0.01	-4.95	0.00
black1	-0.48	0.27	-1.79	0.07
hispan1	0.05	0.36	0.13	0.90
re74	0.00	0.00	3.89	0.00
treat1:hispan1	15.09	722.55	0.02	0.98
treat1:age_c	0.08	0.03	2.86	0.00
treat1:re74	-0.00	0.00	-1.68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

To check whether the AIC model fits that data better than the BIC model, we conduct a Chi-squared test and compared the residual deviance of these two models. 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 improving the fit of our model. We see in the results of our AIC model that the variables black and the interaction between treat and re74 variables have low p-values. Therefore, we created a new model which included these two variables and all the variables in our BIC model. We compared this model to the BIC model using a Chi-squared test. The difference was statistically significant and we decided to keep these two variables in our model. To check if the remaining variables in the AIC model (hispanic 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 we chose not to add hispanic and its interaction with the treat variable to our model.

The results of our model suggest that all the variables are significant except the interaction of treat and re74 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 the 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_centered variables has still not been captured.

We observe that the trend of residuals changes twice against the values of our age_centered variable, convincing us to include the squared and cubic degree polynomials of age_centered variable in our model. We also included the interactions of the treat variable with age_squared and age_cubed variables. We run the regression again with the additional variables to check if this improves our model. 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.

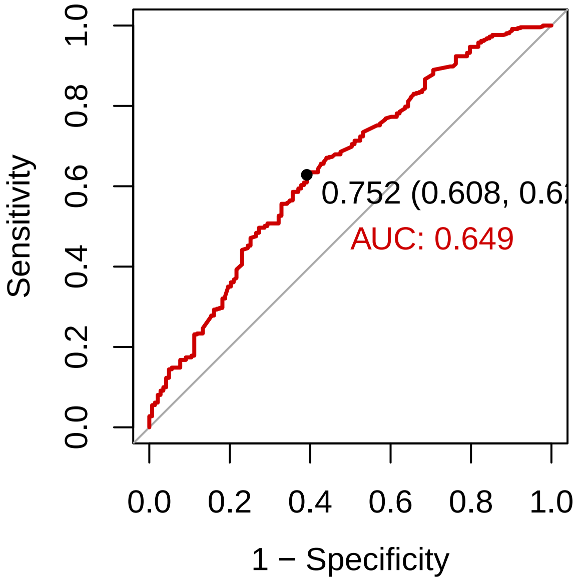




We see that there is not much difference in these plots and the outliers on the left side of the plot are still present. Given that adding the squared and cubic terms of age in our model makes it harder to interpret the standard estimates and does not significantly improve the fit of our model, we decided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

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 treat_{i1} + \beta_2 age(centered)_{i2} + \beta_3 black_{i3} + \beta_4 re74_{i4} + \beta_5 treat{*}age(centered)_{i5} + \beta_6 treat{*}re74_{i6} \tag{1}$$

We now move on to test how well our model performs on predicting outcomes. We use our model to predict outcomes on the train data set available to us and then estimate the accuracy of our model's prediction by measuring the area under the ROC curve which is shown below.



We then use 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 did earned a zero wages in 1978 (11% specificity). Therefore, to obtain a balance between specificity and sensitivity, we classify outcomes based on the probability threshold suggested by the ROC curve, which 0.752. Using this threshold, we are able to achieve sensitivity rate of 63% and specificity rate increases to 60%.

The standard estimates of all the coefficients in our model were statistically significant except the interaction between treat and re74 variable. We exponentiate the standard estimates and their confidence intervals to interpret them on the odds scale. Below are the results of our model.

	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

The intercept of our model tell us that for a non-black person aged 27 years, and unemployed in 1974, the odds of earning positive wages in 1978 without training is 2.77. We see that for an unemployed person in 1974 aged 27, 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 person who did not receive training, a $1 increase in the earnings of 1974, increases the odds of earning positive wages by 0.01%. For a person who did not receive training, a 1 year increase in age decreases the odds of earning positive wages by 5.5%. However,for people who participated in the training, 1 year increase in age leads to an increase in the odds of earning a non-zero wage by 1.9%. We get this value by multiplying the standard estimates (odds scale) of the age variable and interaction variable between age and treat. We also note that the odds of a black person earning positive wages is 0.54 times the odds of a non-black person earning positive wages. Despite having interactions, multicollinearity is not present in our model as the variance inflation factor of each of the variables is below 10. A table indicating the VIFs of all the variables in our model has been added to the appendix.