Team Project 1: Effects of Job Training on Wages

Guillem Amat, Calvin Dong, Jose Moscoso, Varun Prasad, Anshupriya Srivastava

October 4, 2019

Summary

This report investigates the variables that are associated with higher wages and the higher odds of a positive wage. The data analyzed consisted of disadvantaged males who either received or did not receive job training in addition to other demographic variables such as their age, previous wages, race, marital status, and education. A multiple linear regression model was used to analyze if training was associated with an increase in wage. The final valid model had an R^2 of 0.1094 and training was found to not be statistically significant in higher wages. The multiple logistic regression model was used to determine whether training improved the odds of obtaining a positive wage. The final valid model had an accuracy of 58.63% and an AUC of 0.638, indicating relatively poor predictive ability. These models of poor fit can be attributed to factors such as the uneven distribution of the data for the treatment and control groups, the subsetting of the original data into only males, and the heavy presence of zero values for the 1974 wage variable. It is also possible that these models do not best capture the relationship between these predictor variables and higher wages, and another statistical model would be needed.

Introduction

The National Supported Work (NSW) Demonstration was a famous study conducted in the late 1970s to determine whether or not job training for disadvantaged workers improved their wages. Eligible candidates were randomized into groups that would either receive or not receive job training. Additional variables, such as the participant's age, race, marital status, and education were also considered. The analysis presented in the following report investigates whether job training, in addition to these other variables, positiviely impacts the wages of these disadvantaged workers. Both multiple linear and multiple logistic regression models are used to answer questions about improved wages as a result of job training. The linear model addresses whether job training leads to higher wages and the logistic model addresses whether job training increases the odds of having a positive wage. The original data has been filtered to simplify the analysis, and the conditions are specified in the Data section of this report.

Data

The data used for this analysis has been subsetted to consist of only male participants. The treatment group consists of males whose 1974 earnings were used in an NSW study conducted by economist Robert Lalonde. The control group consists of unemployed males whose income in 1975 was below the poverty line. The original dataset was created from a randomized experiment, which would allow us to determine causality. However, subsetting the dataset into males who follow specific criteria only enables us to determine a possible association between job training and improved wages. The data was organized and processed differently for the two different models as described in the following sections. However, in both cases, the values of wages in 1975 (re75) were excluded since these represent wages during the training period.

Linear Model

Initial inspection of the data revealed important insights about the distribution of each variable. Many of the 1978 earnings (re78) were 0. These values were removed since it is unclear whether 0 corresponds to

unemployment, voluntary leave from work, or working but not receiving payment. Furthermore, 429 people did not receive job training while 185 did, indicating an uneven distribution of these two groups of interest. The proportion of Hispanics in the dataset was below 10%, meaning that their influence in the model is likely to be insignificant. Blacks were the second largest racial group in the model, and 156 out of 243 black people received training, making them a highly relevant group for the analysis and likely to affect the final results. 30% of the people in the dataset are between 16 and 20 years old, so it is possible that most of their jobs might be entry-level positions with low salaries. Thus, the impact of the additional training might not be as substantial as expected. There were also a few outliers, both of whom were black people that received treatment with more than 10 years education and no degree. One of them had a huge salary increase of $60,300 and the other with a loss of $25,256. Approximately 10% of the observations were above 43 years old, and 55% of the observations had between 10 and 12 years of education. In a comparison with the distribution of ages, no person older than 27 years old had more than 10 years of education. Most of the people older that 27 years old had between 0 and 3 years of study, which could indicate that most of them work at low skilled jobs with low salary. This latter group should be expected to have improved wages since they previously would not have much education or training that would have increased their likelihood of higher wages. Age and education thus should be important elements in our model.

Exploratory data analysis for the linear model first investigated whether or not the response variable of wages in 1978 (re78) was normally distributed. A histogram showed that even after removing all the values of zero, re78 showed a right skewed distribution. Applying a square root transformation to the data normalized the distribution, thus enabling analysis using multiple linear regression. The histograms of the original and transformed data are shown in the following plot:

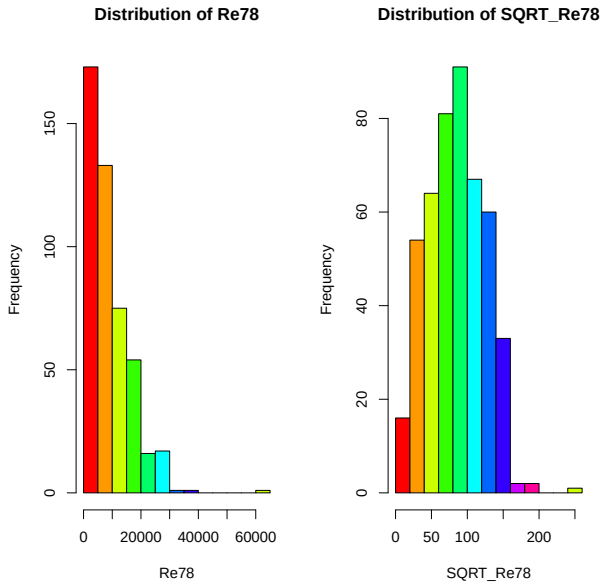


Figure 1: Histograms of Untransformed and Transformed re78

Initially, a linear regression model had been presented without performing this transformation due to the difficult nature of interpreting the coefficients. However, after seeing the class presentations and obtaining feedback, reverting back to the square root transformation was the most valid option. Without a normally distributed response variable, a linear regression model cannot be used to generate meaningful insights.

Plotting the continuous variables of age and education against the square root of re78 showed more linear relationships, further validating the transformation. These variables were then mean-centered in order to improve interpretation and reduce multicollinearity. The wages in 1974 (re74) were also square rooted to maintain consistency with the square rooted re78 values. Boxplots of re78 against categorical variables of

race and marital status were plotted to determine potential interactions. Through analysis of the boxplots, changes in difference by treatment controlled for race (black and hispanic) and marital status (married or not) did not appear to be substantial. Boxplots are shown in Appendix 1.1.

Logistic Model

Exploratory data analysis of the logistic model consisted of the response variable of earnings in 1978 (re78) and the following predictors: treat, age, educ, black, hispanic, married, no degree, and earnings in 1974. The re78 variable was factored so that any positive values were assigned a value of 1 while any non-positive or missing values were assigned a value of 0. For this model, this binary factoring was also applied to the re74 variable in order to analyze whether having a positive wage in 1974 had a positive influence on the odds of a positive wage in 1978. A χ^2 test of deviance was used on each of the categorical variables to determine if there might be any potential association. The results of these tests are shown in the following table:

Training	Black	Hispanic	Married	No Degree	re74
0.7686	0.0201	0.2052	0.5756	0.5053	0.0332

Table 1: P-values of χ^2 Deviance Tests

Based on this initial analysis, the variables that appear to show a significant association with different wages are whether the person was black and whether the person had a job in 1974. All of the other variables have p-values that are greater than the significance level of 0.05. Interestingly, it does not appear that job training is associated with a positive wage, but this can only be confirmed after full analysis.

Binned plots of the continuous variables of age and education against wage were plotted. These plots all primarily showed a linear pattern, though there were some curvature toward the upper end of age. Therefore, no transformation is needed. The continuous variables were then all mean-centered to improve interpretability and eliminate potential effects of multicollinearity. Plots are shown in Appendix 2.1.1.

Boxplots were plotted to consider interactions between the categorical variables. Based on these plots, there appear to be potential interactions between the following pairs: age and treatment; education and hispanic; and education and black. These plots are shown in Appendix 2.1.2.

Models

Linear Regression

Multiple linear regression (MLR) was used to model the data. The general formula for an MLR model is the following:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_p x_{ip} + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

The final model was created with the following predictor variables: treatment, mean-centered age, mean-centered education, whether the individual was black, and the mean-centered square root of the 1974 wage. This model was generated by first incorporating all predictor variables (excluding 1975 wages) into the model and performing stepwise selection using the AIC criterion. Interactions between the following pairs were tested: treatment and age; treatment and race; treatment and education. None of these interactions were significant after performing F-tests. The AIC of the final model was 4754.19, and the results of the model are shown in the following table:

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	85.8997	2.2802	37.67	0.0000
Treatment	8.1381	4.7770	1.70	0.0891
Mean-Centered Age	0.6144	0.2047	3.00	0.0028
Mean-Centered Education	2.3799	0.6854	3.47	0.0006
Race: Black	-7.8755	4.4768	-1.76	0.0792
Mean-Centered $\sqrt{Wage_{1974}}$	0.1591	0.0404	3.94	0.0001

Table 2: Coefficients of Linear Regression Model

The final model thus has the following formula:

$$\sqrt{Wage_{1978}} = \beta_0 + \beta_1 Treatment + \beta_2 Age_{centered} + \beta_3 Educ_{centered} + \beta_4 Black + \sqrt{Wage_{1974}}_{centered} + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

The confidence intervals for the predictor variables are shown in the following table:

	2.5 %	97.5 %
Intercept	81.42	90.38
Treatment	-1.25	17.53
Mean-Centered Age	0.21	1.02
Mean-Centered Education	1.03	3.73
Black	-16.67	0.92
Mean-Centered $\sqrt{Wage_{1974}}$	0.08	0.24

Table 3: Confidence Intervals for Linear Regression Model Parameters

These intervals indicate that we are 95% confident that the true value for each parameter lies within the interval. Based on the p-values and the intervals, only the mean-centered continuous variables are significant. The variables of being black and treatment are not significant in leading to higher wages.

Model validation was performed by assessing the four main assumptions of linearity, normality, equal variance, and independence. Residual plots used for validation are shown in Appendix 1.2. Linearity was assessed by plotting the continuous predictor variables of the mean-centered age, education, and square root of 1974 wage against the residuals. None of these plots showed a discernible pattern, although there is a concentration of points on the left side of the residual plot for the square root of 1974. Therefore, linearity is met. A normal q-q plot shows some deviation at both extremes of the 45 degree line, but most of the the points fall near or over the line of identity, thus validating normality. A plot of the fitted values against the residuals shows an even and random distribution about 0, thus validating both equal variance and independence. Finally, a plot of the residuals against the leverage does not show any outliers. Multicollinearity was also assessed and all of the coefficients were near 1, indicating almost no correlation between the variables. The R^2 of the model was 0.1094 with a standard residual error of 37.32. Thus while the linear model is valid, it is clearly not a strong fit for the data.

Logistic Regression

Because the predictor variable of positive wage is represented in a binary form (0 for a positive wage and 1 for a non-positive wage), a logistic regression model with multiple predictors was used to fit the data. This model, also known as the logit function, is represented by the following equation:

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

where $y_i|x_i \sim$ Bernouilli(π_i).

The logit function calculates the log-odds of the response variable due to changes in the predictor variables. Exponentiating the resulting coefficients will give the odds of a response variable depending on changes in the predictor variables. Prior to model selection, the continuous variable of age was centered to better explain the log-odds at the averages of those respective variables. The full logistic model contained the following predictor variables: mean-centered age, black, treatment, and an interaction term between treatment and age. To generate the model, forward, backward, and stepwise selection were conducted with AIC as the test criterion. All three methods resulted in a model with mean-centered age, black, treatment, and re74 as significant. χ^2 Tests of Deviance were conducted between this selected model and models with the pairs of interactions mentioned in the Data section. After additional testing of interactions, only the variables of treatment and the interaction term between treatment and age were added to the final model due to their statistical significance. The final model had an AIC of 648.475. The coefficients of the final model are shown in the following table:

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.8763	0.2294	3.82	0.0001
Mean-centered Age	-0.0441	0.0105	-4.19	0.0000
Treatment	0.8623	0.3244	2.66	0.0079
Race: Black	-0.6278	0.2504	-2.51	0.0122
Positive 1974 Wage	0.7968	0.2634	3.03	0.0025
Treatment and Mean-centered Age Interaction	0.0634	0.0275	2.30	0.0212
Treatment and Positive 1974 Wage Interaction	-0.8868	0.4599	-1.93	0.0538

Table 4: Coefficients of Log-Odds of Logistic Regression Model

The above coefficients indicate how the log-odds of receiving a positive wage change with each variable, assuming all other variables are held constant.

Confidence intervals for change in odds based on the model's parameters are shown in the following table:

	2.5 %	97.5 %
Intercept	0.44	1.34
Mean-centered Age	-0.06	-0.02
Treatment	0.23	1.51
Race: Black	-1.12	-0.14
Positive 1974 Wage	0.28	1.31
Mean-centered Age and Treatment Interaction	0.01	0.12
Positive 1974 Wage and Treatment Interaction	-1.78	0.03

Table 5: Confidence Intervals of Odds for Parameters of Logistic Regression Model

These intervals indicate that we are 95% confident that the true change in odds of receiving a positive wage are within each interval. Note that the confidence interval for the interaction between a positive 1974 wage and treatment contains zero, so this variable is not significant in predicting the change in odds of a positive wage. The intervals of all other variables exclude zero, so they are all significant.

The model thus has the final following formula:

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 Age_{centered} + \beta_2 Training + \beta_3 Black + \beta_4 Wage_{1974} + \beta_5 Age_{centered} : Training + \beta_6 Wage_{1974} : Training$$

where $y_i|x_i \sim$ Bernouilli(π_i) and π_i is the probability of a positive wage in 1978.

Model validation was conducted using binned residual plots. The plot of the fitted values against the residuals showed a random pattern with only one point outside the confidence band. A plot of the continuous variable

of mean-centered age against the residuals also shows a generally random pattern with only a few points outside the confidence bands.

A confusion matrix is also used to further validate the model. Due to the uneven distribution of those who received training and those who did not, the cutoff value was set as the mean of the positive wage variable, calculated to be 0.767. The accuracy, specificity, and sensitivity of the confusion matrix are shown in the following table:

Cutoff	0.7671
Accuracy	0.5863
Sensitivity	0.5839
Specificity	0.5944

Table 6: Results of Confusion Matrix

The ROC curve for the model has an AUC of 0.638 and is shown in the following plot:

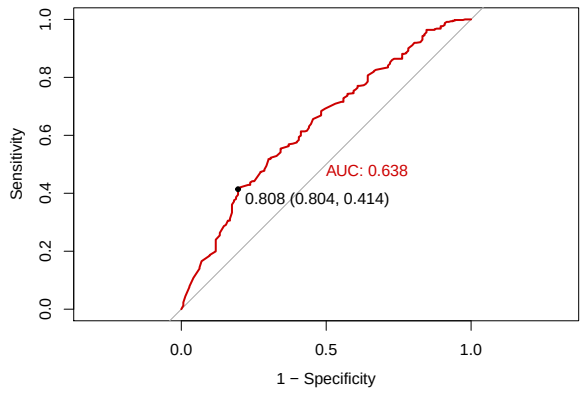


Figure 2: ROC Curve

Overall, the model is valid and can be used to generate insights about the data.

Results

Linear Regression

Because a square root transformation was applied to the response variable of re78, the following approach was used to interpret the coefficients:

$$\widehat{\sqrt{y}} = \beta_0 + \beta_1 X_1$$
$$\widehat{y} = (\beta_0 + \beta_1 X_1)^2$$
$$\frac{\mathrm{d}\widehat{y}}{\mathrm{d}x_1} = 2\beta_1(\beta_0 + \beta_1 X_1) = 2\beta_1 \widehat{y}$$

Effectively, the response variable is squared, and then the derivative is taken. Each predictor variable is kept constant other than the variable of interest in order to determine its effect on the response variable. From this approach, the coefficients of the significant varibles can be interpreted more easily. For the intercept, when every categorical variable is at baseline (0), and the continuous variables each of their mean values,

the wage in 1978 is the square of 85.89, i.e. $7,378. For the beta coefficient of age, for each yearly increase in age, the associated change in re78 is 2*0.614 times the current wage. Following the same estimation, for each additional year of education, the change in re78 is 2*2.379 times the current wage. Finally, for each additional dollar in 1974, the wage in 1978 increased by 2*0.025 times the current wage.

Based on the results in Tables 2 and 3, several important associations can be determined. The p-values and confidence values of the predictor variables show that treatment is not significant in leading to increased wages. Only the mean-centered continous variables of age, education, and the square root of 1974 wages are significant. Marital status, and degree status do not have any significant effects on increased wages and were excluded from the model following the stepwise selection process. The variable of being black, despite being statistically insignificant, was kept since a large proportion of the sample was black.

Logistic Regression

Based on the results in Tables 4 and 5, several important associations can be determined. Based on the coefficients, p-values, and confidence intervals, it is clear that mean-centered age, receiving training, whether or not an individual is black, whether or not an individual has a job in 1974, and the interaction between treatment and age are all significant. The intercept indicates that for people of mean age who were not black, did not receive training, and did not have a job in 1974, the odds of a positive wage is 140.2%. With all other variables held constant, every yearly increase in age decreases the odds a positive wage by 4.3%. With all other variables constant, compared to people who did not receive the training, the odds of a positive wage increase by 136.9% for those who received the training. With all other variables held constant, compared to people who are not black, the odds of a positive wage decrease by 52.9% for those who are black. With all other variables held constant, compared to those who did not have a job in 1974, the odds of a positive wage increase by 121.8%. The coefficient for the interaction term indicates that, with all other variables held constant, for those who received training, each increase in age increases the odds of a positive wage by 6.2%. Numerous other interactions, such as education vs. Hispanic and education vs. black, were tested using a change-in-deviance test, and none of these were found to be significant. Therefore, the only demographic factors that affected the odds of a positive wage were the person's age and whether or not he was black. Interestingly, the interaction between positive wage in 1974 and receiving training is not significant since its p-value is 0.053, though it is still very close to 0.05. Adding this interaction term did cause treatment to be significant, so it was included in the final model.

Conclusion

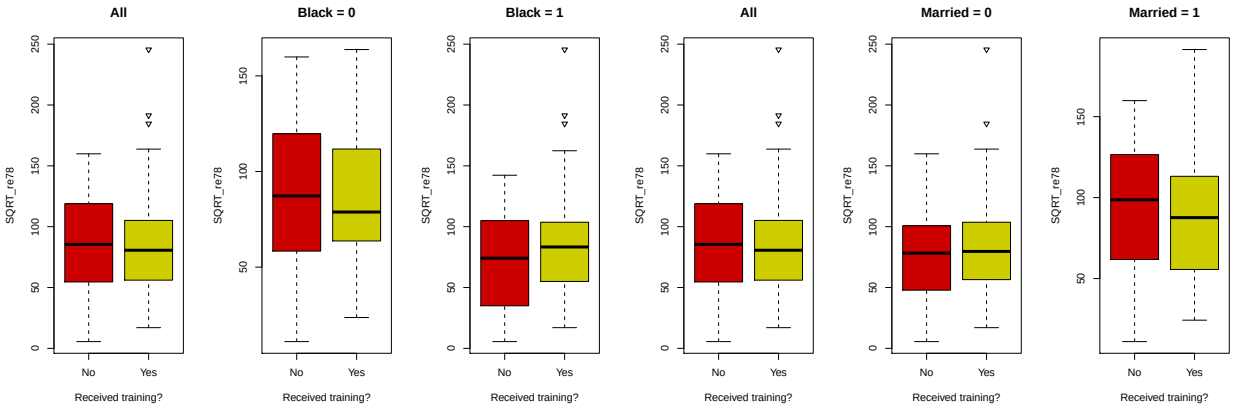
Overall, both models were valid and showed an association between improved wages and predictor variables such as the mean-centered age, mean-centered education, and wages of 1974. The linear model showed that training is not significant in increasing wages while the logistic model showed that the treatment does improve the odds of obtaining a job. Despite the models' validity, the R^2 value of 0.1094 and the accuracy of 58.63% for the linear and logistic models, respectively, indicates that they are poor fits for the data. This finding can be attributed to several factors. First, the sample analyzed was only a subset of the original data set and consisted of only disadvantaged males. Therefore, while the models can generate some insight about the data ananlyzed, they cannot make meaningful predictions about the general population since women and non-disadvantaged people are among those not in the dataset. To more accurately assess the impact of training on wages, a larger dataset containing those specific types of people would be needed. As mentioned earlier, the data itself was also unevenly distributed, leading to skewness of the data and the need for a square root transformation. The proportions of those who received treatment and those who did not was not close to equal, thus making predictions more difficult. There were also many zero values for the re78 data, and having more context for what these values meant could improve classification of the variable and thus the model. It is also unclear whether the training was customized for different jobs or if a general training process was given to all treated individuals. For some people, a nonspecific training would not have been beneficial and could even have hindered their progress. Finally, it is possible that neither linear nor logistic

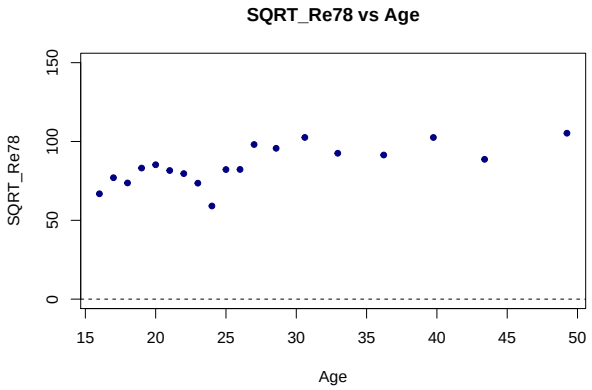
models are the most appropriate for analyzing this dataset, so a completely different model might be needed to truly assess if job training improved the wages of disadvantaged workers.

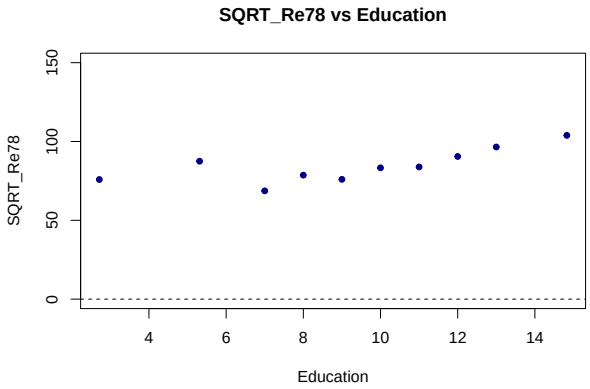
Appendices

Appendix 1: Linear Regression Model

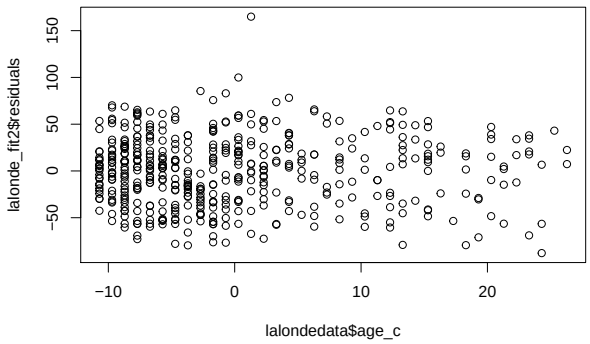
Appendix 1.1: EDA

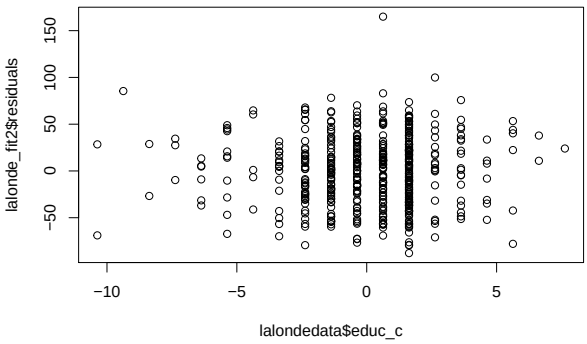


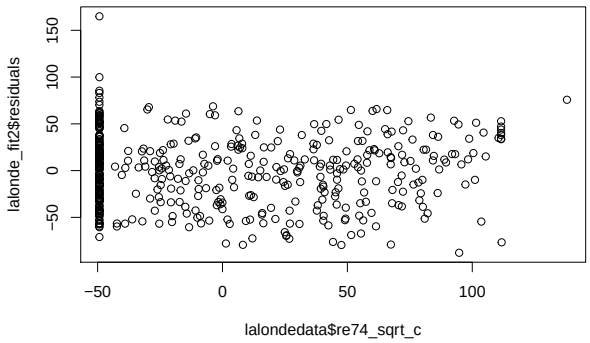


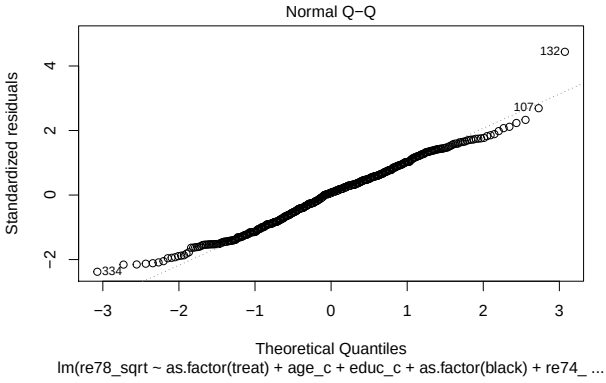


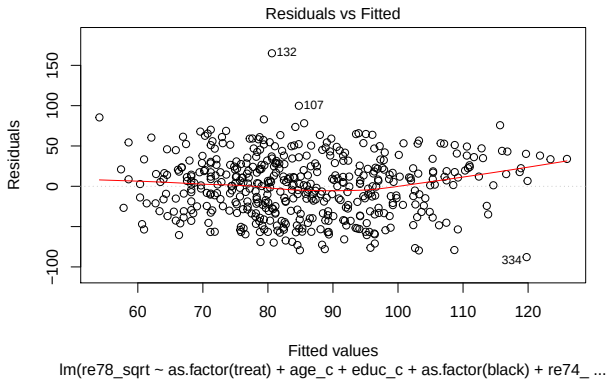
Appendix 1.2: Model Validation

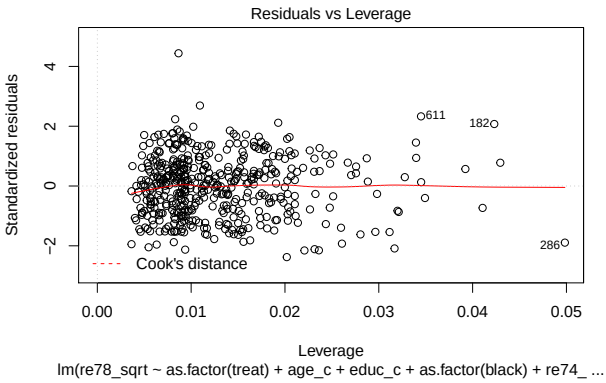








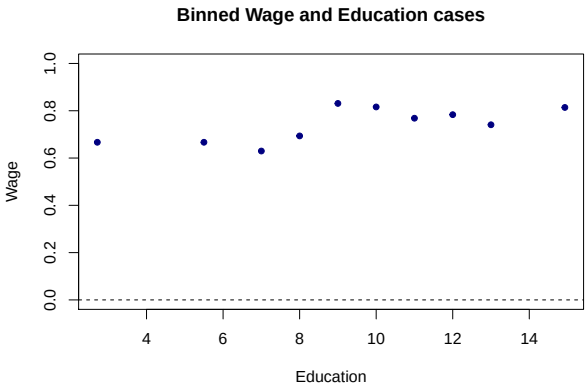




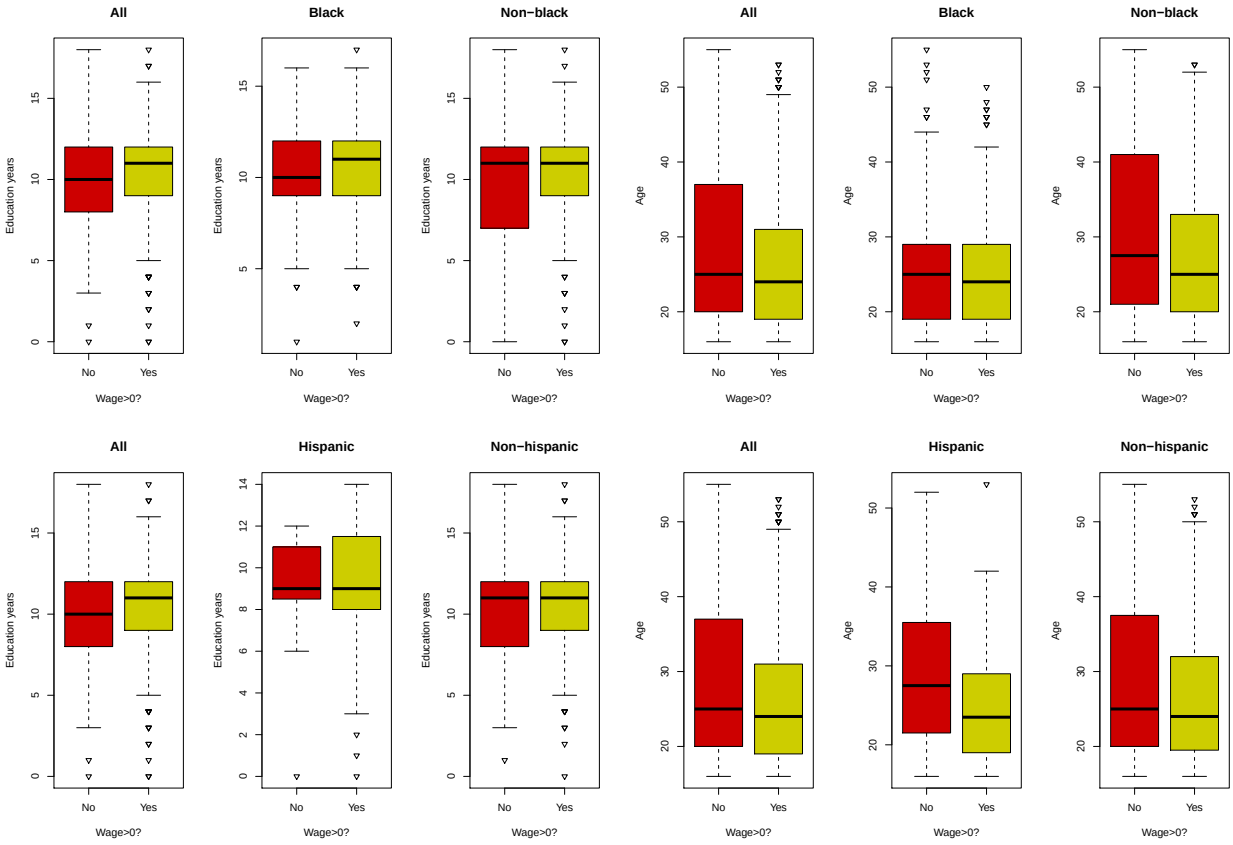
Appendix 2: Logistic Regression Model

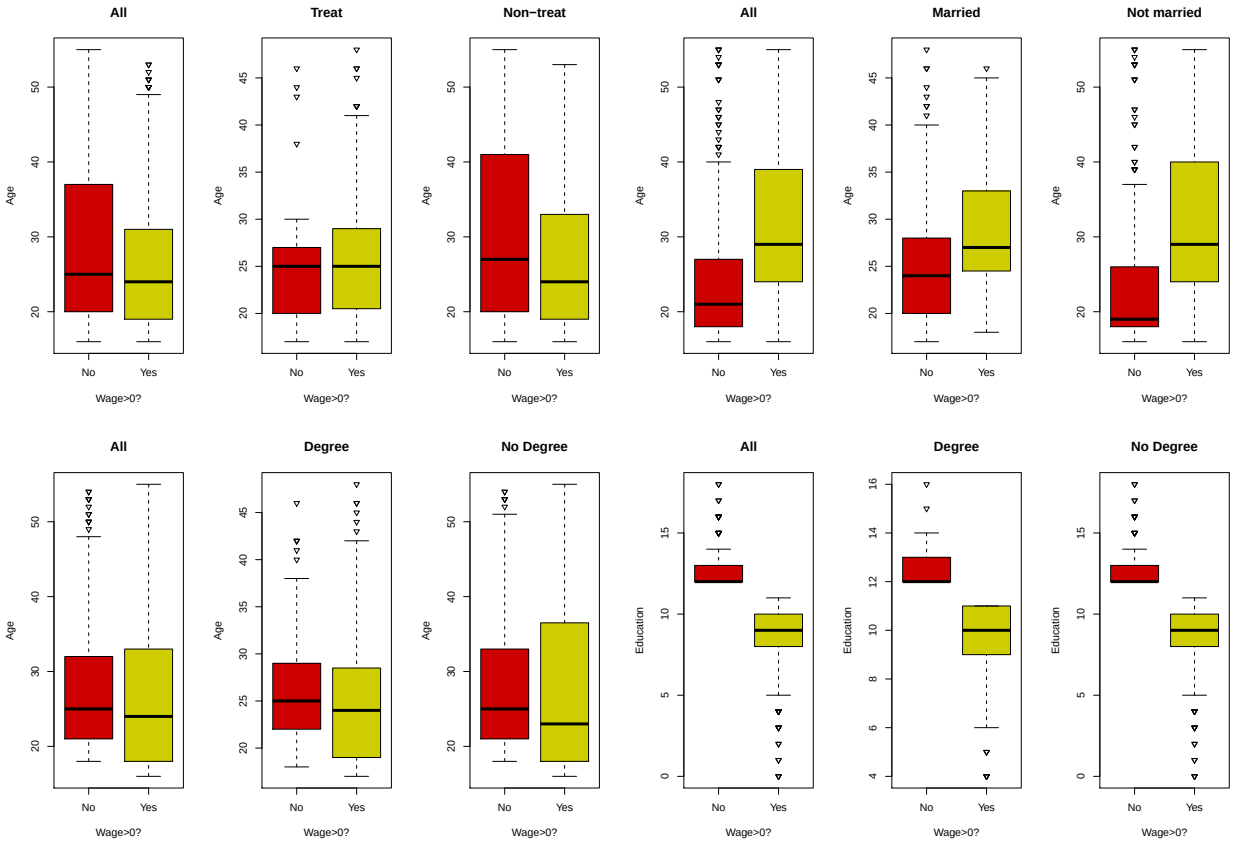
Appendix 2.1: EDA

Appendix 2.1.1: Continuous Variables

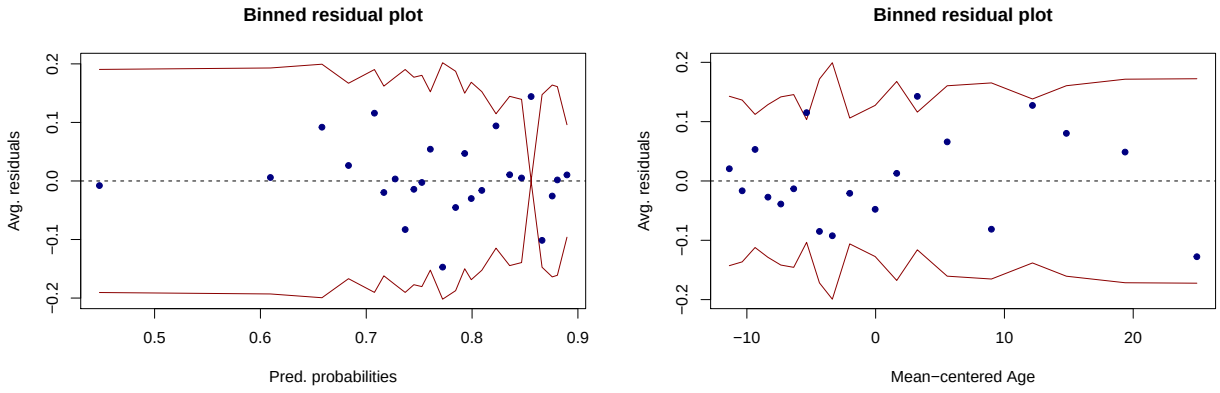


Appendix 2.1.2: Interaction Testing





Appendix 2.2: Model Validation



Appendix 3: Code

```
knitr::opts_chunk$set(out.width="50%",out.height="50%",
                      echo=FALSE, warning=FALSE, message=FALSE,
                      fig.pos = "H")
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(leaps)
library(lattice)
library(rms)
library(gridExtra)
library(arm)
library(e1071)
library(caret)
library(pROC)
library(xtable)
library(tidyverse)
## Import Data
data <- read.table("lalondedata.txt", sep=",", header=TRUE)
# Same data set but used since models were coded differently (dataframes, etc.)
lalonde <- read.table("lalondedata.txt", sep = ",", header = TRUE)
summary(data)
lalonde$re78 <- as.numeric(lalonde$re78)

str(lalonde)
colnames(lalonde)

###### Square root wage variables
lalondedata <- lalonde[lalonde$re78 != 0, ]
lalondedata$re78_sqrt <- sqrt(lalondedata$re78)
lalondedata$re74_sqrt <- sqrt(lalondedata$re74)
lalondedata$re75_sqrt <- sqrt(lalondedata$re75)
lalondedata$age_c <- lalondedata$age - mean(lalondedata$age)
lalondedata$educ_c <- lalondedata$educ - mean(lalondedata$educ)
lalondedata$re74_sqrt_c <- lalondedata$re74_sqrt - mean(lalondedata$re74_sqrt)

par(mfrow=c(1,2))
hist(lalondedata$re78,xlab="Re78",main="Distribution of Re78",col=rainbow(10))
hist(lalondedata$re78_sqrt,xlab="SQRT_Re78",main="Distribution of SQRT_Re78",col=rainbow(10))

#### Understanding the distribution of variables among the categorical variables

table(lalonde$treat)

table(lalonde$age, lalonde$educ)

table(lalonde$hispan)

table(lalonde$black)

table(lalonde$hispan, lalonde$treat)
## 11 out of 61 hispans received treatment.
```

```r
table(lalonde$black, lalonde$treat)
## 156 out of 243 black people received treatment.

table(lalonde$educ, lalonde$treat)
## Most people had between 9 and 12 years of education

table(lalonde$married, lalonde$treat)

table(lalonde$nodegree, lalonde$treat)

#Distribution of observations in age
aggregate(cbind(count = X) ~ age,
          data = lalonde,
          FUN = function(x){NROW(x)})

#Distribution of observations in education
aggregate(cbind(count = X) ~ educ,
          data = lalonde,
          FUN = function(x){NROW(x)})
## Data for logistic model
# Create binary wage for 1978
data$wage.grp[data$re78>0] <- 1
data$wage.grp[is.na(data$wage.grp)] <- 0

# Create binary wage for 1974
data$wage.grp.pre[data$re74>0] <- 1
data$wage.grp.pre[is.na(data$wage.grp.pre)] <- 0

# Create factor for binary wage 1974
data$wage.f <- factor(data$wage.grp, levels=c(0,1), labels = c("N","Y"))
data$wage.f.pre <- factor(data$wage.grp.pre, levels=c(0,1), labels = c("N","Y"))

# Factor other variables
data$treat.f <- factor(data$treat, levels=c(0,1), labels = c("N","Y"))
data$black.f <- factor(data$black, levels=c(0,1), labels = c("N","Y"))
data$hispan.f <- factor(data$hispan, levels=c(0,1), labels = c("N","Y"))
data$married.f <- factor(data$married, levels=c(0,1), labels = c("N","Y"))
data$nodegree.f <- factor(data$nodegree, levels=c(0,1), labels = c("N","Y"))

# Mean center age
data$age.c <- data$age - mean(data$age)

# Mean center education
data$educ.c <- data$educ - mean(data$educ)

par(mfrow=c(1,2))
hist(lalondedata$re78,xlab="Re78",main="Distribution of Re78",col=rainbow(10))
hist(lalondedata$re78_sqrt,xlab="SQRT_Re78",
     main="Distribution of SQRT_Re78",col=rainbow(10))
#EDA

### SQRT_78 vs treat controlled by race(black)
par(mfrow=c(1,3))
```

```r
boxplot(re78_sqrt~treat,data=lalondedata,
        ylab="SQRT_re78",pch=25,xaxt='n',
        xlab="Received training?",
        col=c("red3","yellow3"),cex = 0.85,main ="All")
axis(1,at=c(1,2),labels=c("No","Yes"))

boxplot(re78_sqrt~treat,data=lalondedata,
        subset= black==0, ylab="SQRT_re78",pch=25,xaxt='n',
        xlab="Received training?",
        col=c("red3","yellow3"),cex = 0.85,main ="Black = 0")
axis(1,at=c(1,2),labels=c("No","Yes"))

boxplot(re78_sqrt~treat,data=lalondedata,
        subset= black==1,
        ylab="SQRT_re78",pch=25,xaxt='n',
        xlab="Received training?",
        col=c("red3","yellow3"),cex = 0.85,main ="Black = 1")
axis(1,at=c(1,2),labels=c("No","Yes"))

### SQRT_78 vs treat controlled by race(married)
par(mfrow=c(1,3))
boxplot(re78_sqrt~treat,data=lalondedata,
        ylab="SQRT_re78",pch=25,xaxt='n',
        xlab="Received training?",
        col=c("red3","yellow3"),cex = 0.85,main ="All")
axis(1,at=c(1,2),labels=c("No","Yes"))

boxplot(re78_sqrt~treat,data=lalondedata,subset= married==0,
        ylab="SQRT_re78",pch=25,xaxt='n',
        xlab="Received training?",
        col=c("red3","yellow3"),cex = 0.85,main ="Married = 0")
axis(1,at=c(1,2),labels=c("No","Yes"))

boxplot(re78_sqrt~treat,data=lalondedata,subset= married==1,
        ylab="SQRT_re78",pch=25,xaxt='n',
        xlab="Received training?",
        col=c("red3","yellow3"),cex = 0.85,main ="Married = 1")
axis(1,at=c(1,2),labels=c("No","Yes"))
## Plot continuous predictors
# Binnedplots of age versus re78
binnedplot(y=lalondedata$re78_sqrt,lalondedata$age,
           xlab="Age",ylim=c(0,150),col.pts="navy",
           ylab ="SQRT_Re78",main="SQRT_Re78 vs Age", col.int="white")

#Binnedplots of education versus re78
binnedplot(y=lalondedata$re78_sqrt,lalondedata$educ,
           xlab="Education",ylim=c(0,150),col.pts="navy",
           ylab ="SQRT_Re78",main="SQRT_Re78 vs Education", col.int="white")
###### Significance Analysis of treatment as a predictor of change in salary.

res <- t.test(re78_sqrt ~ treat, data = lalondedata)
res
## Binnedplot for continuous variable vs. response variable
```

```
# Age
binnedplot(y=data$wage.grp,data$age,
           xlab="Age",ylim=c(0,1),col.pts="navy",
           ylab ="Wage",
           main="Binned Wage and Age cases",col.int="white")
# Education
binnedplot(y=data$wage.grp,data$educ,
           xlab="Education",ylim=c(0,1),col.pts="navy",
           ylab ="Wage",
           main="Binned Wage and Education cases",col.int="white")
## 2 by 2 tables and chisq test for categorical variable vs response variable
# Treatment
table(data[,c("wage.f","treat.f")])/sum(table(data[,c("wage.f","treat.f")]))
chisq.test(table(data$wage.f,data$treat.f))

# Black
table(data[,c("wage.f","black.f")])/sum(table(data[,c("wage.f","black.f")]))
chisq.test(table(data$wage.f,data$black.f))

# Hispanic
table(data[,c("wage.f","hispan.f")])/sum(table(data[,c("wage.f","hispan.f")]))
chisq.test(table(data$wage.f,data$hispan.f))

# Married
table(data[,c("wage.f","married.f")])/sum(table(data[,c("wage.f","married.f")]))
chisq.test(table(data$wage.f,data$married.f))

# No Degree
table(data[,c("wage.f","nodegree.f")])/sum(table(data[,c("wage.f","nodegree.f")]))
chisq.test(table(data$wage.f,data$nodegree.f))

# Wage 1974
table(data[,c("wage.f","wage.f.pre")])/sum(table(data[,c("wage.f","wage.f.pre")]))
chisq.test(table(data$wage.f,data$wage.f.pre))
##Education years vs. wage group by black
par(mfrow=c(1,3))
boxplot(educ~wage.f,data=data,ylab="Education years",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="All")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(educ~wage.f,data=data, subset= black==1, ylab="Education years",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Black")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(educ~wage.f,data=data, subset= black==0, ylab="Education years",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Non-black")
axis(1,at=c(1,2),labels=c("No","Yes"))


## age vs. wage group by black
par(mfrow=c(1,3))
boxplot(age~wage.f,data=data,ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="All")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(age~wage.f,data=data, subset= black==1, ylab="Age",pch=25,xaxt='n',
```

```
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Black")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(age~wage.f,data=data, subset= black==0, ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Non-black")
axis(1,at=c(1,2),labels=c("No","Yes"))



## education years vs. wage group by hispanic
par(mfrow=c(1,3))
boxplot(educ~wage.f,data=data,ylab="Education years",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="All")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(educ~wage.f,data=data, subset= hispan==1, ylab="Education years",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Hispanic")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(educ~wage.f,data=data, subset= hispan==0,
        ylab="Education years",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),
        cex = 0.85,main ="Non-hispanic")
  axis(1,at=c(1,2),labels=c("No","Yes"))


##age vs. wage group by hispanic
par(mfrow=c(1,3))
boxplot(age~wage.f,data=data,ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="All")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(age~wage.f,data=data, subset= hispan==1, ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Hispanic")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(age~wage.f,data=data, subset= hispan==0, ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Non-hispanic")
axis(1,at=c(1,2),labels=c("No","Yes"))

## Age vs. wage group by treat
par(mfrow=c(1,3))
boxplot(age~wage.f,data=data,ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="All")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(age~wage.f,data=data, subset= treat==1, ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Treat")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(age~wage.f,data=data, subset= treat==0, ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Non-treat")
axis(1,at=c(1,2),labels=c("No","Yes"))

## Age vs. wage by married
par(mfrow=c(1,3))
boxplot(age~married.f,data=data,ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="All")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(age~married.f,data=data, subset= treat==1, ylab="Age",pch=25,xaxt='n',
```

```
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Married")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(age~married.f,data=data, subset= treat==0, ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Not married")
axis(1,at=c(1,2),labels=c("No","Yes"))

# Age vs. wage by degree
par(mfrow=c(1,3))
boxplot(age~nodegree.f,data=data,ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="All")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(age~nodegree.f,data=data, subset= treat==1, ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Degree")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(age~nodegree.f,data=data, subset= treat==0, ylab="Age",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="No Degree")
axis(1,at=c(1,2),labels=c("No","Yes"))

# Education vs. wage by degree
par(mfrow=c(1,3))
boxplot(educ~nodegree.f,data=data,ylab="Education",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="All")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(educ~nodegree.f,data=data, subset= treat==1,
        ylab="Education",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="Degree")
axis(1,at=c(1,2),labels=c("No","Yes"))
boxplot(educ~nodegree.f,data=data, subset= treat==0,
        ylab="Education",pch=25,xaxt='n',
        xlab="Wage>0?",col=c("red3","yellow3"),cex = 0.85,main ="No Degree")
axis(1,at=c(1,2),labels=c("No","Yes"))

lalonde_fit <- lm(re78_sqrt ~ as.factor(treat) + age_c + educ_c
                  + as.factor(black) + as.factor(hispan)
                  + as.factor(married) + as.factor(nodegree)
                  + re74_sqrt_c, data = lalondedata)

summary(lalonde_fit)

# Stepwise AIC to eliminate insignificant variables
step(lalonde_fit, scope=formula(lalonde_fit),direction="both",
     trace=0,k = 2)

# Model using AIC significant variables
lalonde_fit2 <- lm(re78_sqrt ~ as.factor(treat) + age_c + educ_c
                   + as.factor(black) + re74_sqrt_c,
                   data = lalondedata)
summary(lalonde_fit2)

###### Check for interaction

#### treat and age
#It is still not significant 0.7472
```

```r
lalonde_fit2 <- lm(re78_sqrt ~ as.factor(treat) + age_c + educ_c
                   + as.factor(black) + re74_sqrt_c,
                   data = lalondedata)


lalonde_fit_test <- lm(re78_sqrt ~ as.factor(treat) + age_c
                       + as.factor(treat):age_c + educ_c
                       + as.factor(black) + re74_sqrt_c,
                       data = lalondedata)

anova(lalonde_fit2, lalonde_fit_test)


#### treat and education
#It is still not significant 0.5894
lalonde_fit2 <- lm(re78_sqrt ~ as.factor(treat) + age_c + educ_c
                   + as.factor(black) + re74_sqrt_c,
                   data = lalondedata)


lalonde_fit_test <- lm(re78_sqrt ~ as.factor(treat) + age_c
                       + as.factor(treat):educ_c + educ_c
                       + as.factor(black) + re74_sqrt_c,
                       data = lalondedata)

anova(lalonde_fit2, lalonde_fit_test)



#### treat and black
#It is still not significant 0.3481
lalonde_fit2 <- lm(re78_sqrt ~ as.factor(treat) + age_c + educ_c
                   + as.factor(black) + re74_sqrt_c,
                   data = lalondedata)


lalonde_fit_test <- lm(re78_sqrt ~ as.factor(treat) + age_c
                       + as.factor(treat):black + educ_c +
                         as.factor(black) + re74_sqrt_c,
                       data = lalondedata)

anova(lalonde_fit2, lalonde_fit_test)

# Multicollinearity
vif(lalonde_fit2)
###Plot of residuals
# Linearity
plot(lalonde_fit2$residuals~lalondedata$age_c)
plot(lalonde_fit2$residuals~lalondedata$educ_c)
plot(lalonde_fit2$residuals~lalondedata$re74_sqrt_c)

# Normality
plot(lalonde_fit2, which = 2)
```

```r
# Variance/Independence
plot(lalonde_fit2, which = 1)

# Outlier
plot(lalonde_fit2, which = 5)
## Model selection using AIC
#
NullModel <- glm(wage.f~1,data=data,family=binomial)
FullModel <- glm(wage.f~age.c+educ.c+treat.f+black.f+hispan.f+married.f
                 +nodegree.f+wage.f.pre,data=data,family=binomial)

# Backward selection
model_backward <- step(FullModel, scope = formula(NullModel),direction="backward", trace = 0)

# Stepwise selection
model_stepwise <- step(NullModel, scope = formula(FullModel),direction="both", trace = 0)

# Forward selection
model_forward <- step(NullModel, scope = formula(FullModel), direction = "forward", trace = 0)
# Model without wage1974
model_1 <- glm(wage.f~age.c + treat.f + black.f + age.c:treat.f,
                            data = data, family = binomial)

# Test interaction between age and treat
model_test_age_treat <- glm(wage.f~age.c + treat.f + black.f + wage.f.pre
                              + age.c:treat.f,
                            data = data, family = binomial)
anova(model_stepwise, model_test_age_treat, test = "Chisq") # Significant

# Test other interaction pairs: education vs. black, hispanic
model_test_hisp_edu <- glm(wage.f~age.c + treat.f + black.f + age.c:treat.f +
                             hispan.f + educ + hispan.f:educ, data = data,
                           family = binomial)
anova(model_test_age_treat, model_test_hisp_edu, test = "Chisq")
# Not significant

model_test_black_edu <- glm(wage.f~age.c + treat.f + black.f + age.c:treat.f +
                              educ + black.f:treat.f, data = data,
                            family = binomial)
anova(model_test_age_treat, model_test_black_edu, test = "Chisq")
# Not significant

# Interaction between treat and wage1974
model_test_1974_treat <- glm(wage.f~age.c + treat.f + black.f + wage.f.pre
                              + age.c:treat.f + wage.f.pre:treat.f,
                            data = data, family = binomial)
anova(model_test_age_treat, model_test_1974_treat, test = "Chisq")
# Almost significant: treatment term becomes significant after doing summary

# Final model
main_model <- model_test_1974_treat
## Logistic Model Validation
# Binned residual plots
```

```r
binnedplot(fitted(main_model),residuals(main_model,"resp"),
           xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot",col.pts="navy")

binnedplot(data$age.c,residuals(main_model,"resp"),
           xlab="Mean-centered Age",
           col.int="red4",
           ylab="Avg. residuals",
           main="Binned residual plot",col.pts="navy")
# Confusion matrix

conf_mat <- confusionMatrix(as.factor(ifelse(fitted(main_model) >=
                                               mean(data$wage.grp), "Y","N")),
                            data$wage.f,positive = "Y")

conf_mat$table
conf_mat$overall["Accuracy"]
conf_mat$byClass[c("Sensitivity","Specificity")]
# ROC curve
invisible(roc(data$wage.f,fitted(main_model),
              plot=T,print.thres="best",legacy.axes=T,
              print.auc =T,col="red3",
              fig_caption = "ROC Curve of Logistic Model"))
```