Team Project I
Effects of Job Training on Wages

Tien Yu Liu, Christy Hu, Ronald Nhondova, Srishti Saha, and Altamash Rafiq

Introduction

Fifty years ago, one of the most famous experiments on the evaluation of public policy programs in the United States was conducted. In this National Supported Work (NSW) Demonstration, researchers were interested in assessing whether or not job training for disadvantaged workers had an impact on their wages. This report considers a subsection of the data from this study to explore similar questions to the ones the researchers who conducted the study considered:

- Is there evidence that workers who receive job training tend to earn higher wages 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?
- Does the effect of the treatment differ by race/demographics?
- What are other associations worth noting with the income in 1978.

A multiple linear regression model was developed to map the data and interpretted to posit answers to the afore mentioned questions of interest.

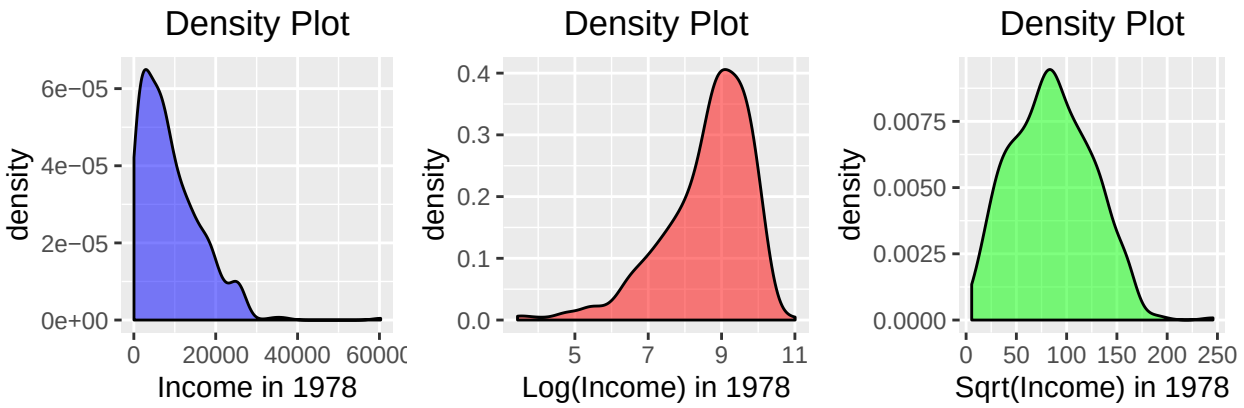
Data

Data Pre-Processing

The data used in this analysis is of 471 male participants with the treatment group including those whose 1974 earnings can be obtained. The control group includes unemployed males in 1976 whose income in 1975 was below the poverty level. As the data for the year 1975 is simply data for an intermediate stage of this study and also contains cases where participants were paid to join the study, it was determined that the 1975 data should not be used in the analysis as it may hinder the fitting of a stronger model. In addition, 143 observations that have zeros values in 1978 (*re78*) are removed.

Exploratory Data Analysis

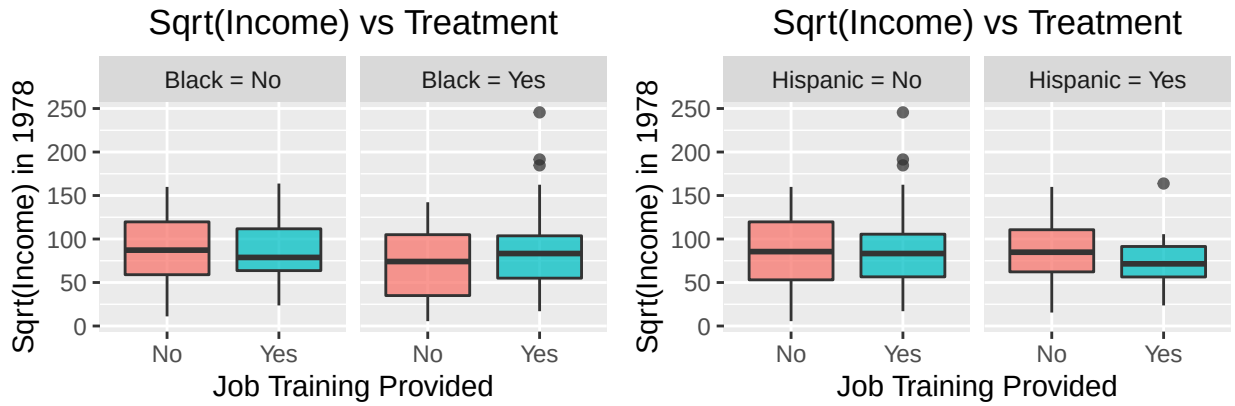
Before any analysis was conducted, a density plot of the response variable, the income in 1978 of the participants, was produced to assess if the assumption of normality that is necessary for linear regression is satisfied. As can be seen from the plots below, a significant left skew in the data was observed, created by the fact that the majority of the incomes in 1978 were closer to zero than otherwise, followed by a thick tail of higher values. Seeing this, it was determined that a transformation of the response variable will result in an increase in its normality. Log and square root transformations were explored to rectify the data skew with the square root transformation leading to a more normally distributed (and unskewed) response varaible. As such, it was selected as the preferred response variable for the modeling.



In the process of exploring the data, a large number of plots and summary statistics were generated to identify associations between the variables and make strong judgements about which predictors to include in the final model. Additionally, these plots and statistics were used to identify potential concerns within the data that may impact the analysis.

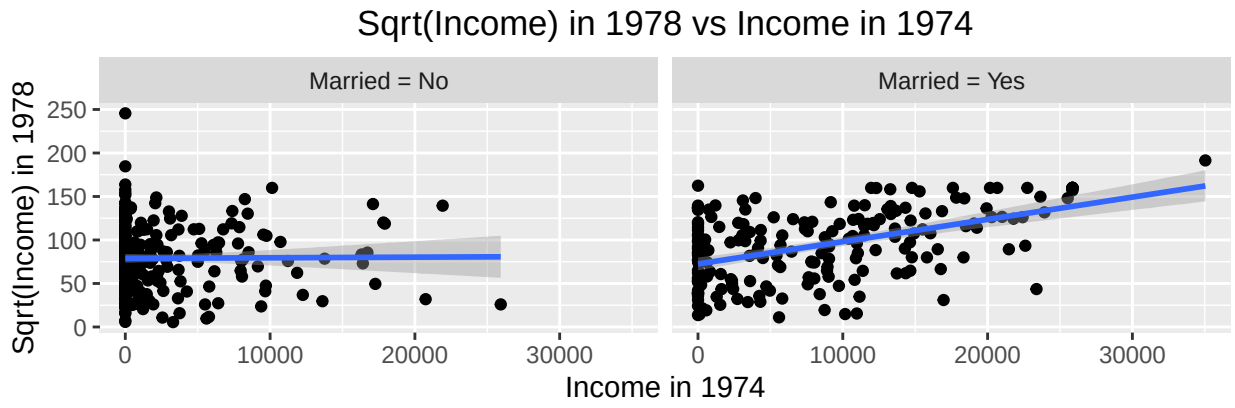
At this stage, the indicator variables for whether or not the person was given the job training, whether or not the person was black, whether or not the person was hispanic, whether or not the person has a degree, and whether or not the person was married were identified as poor predictors of the response variable. The boxplots of these variables against the square root of income in 1978 did not indicated much difference by category. Associations between square rooted income in 1978 and income in 1974, age, and education were identified in this exploratory stage.

Additionally, plots were made to assess possible interaction relationships between variables. In particular, the boxplots for the interactions between the indiactor for black and the indicator for treatment as well as the indiactor for hispanic and the indicator for treatment were closely scrutinized for evidence that the association between square rooted income in 1978 and treatment differs by these races:



As per these graphs, there appears to be a slight differnce in the boxplots for treatment versus square rooted income in 1978 for both races but the difference is not large enough to convince us that the difference would likely be characterized as significant in the modeling process.

Additionally, a relationship of interest was discovered between square rooted income in 1978 and income in 1974 segmented by whether or not the subject was married:

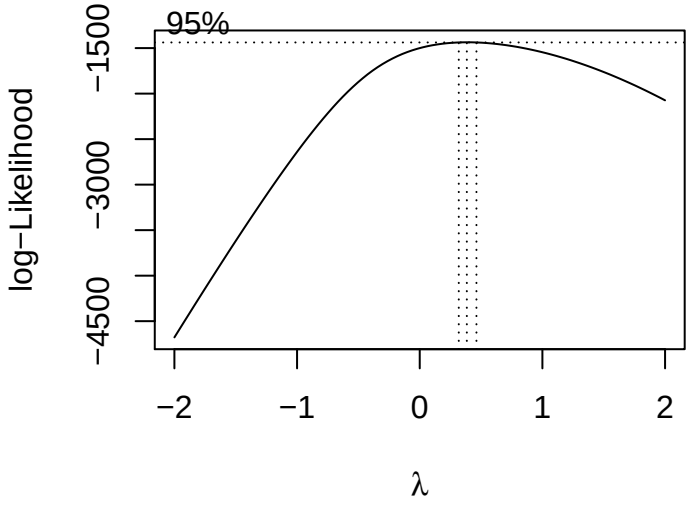


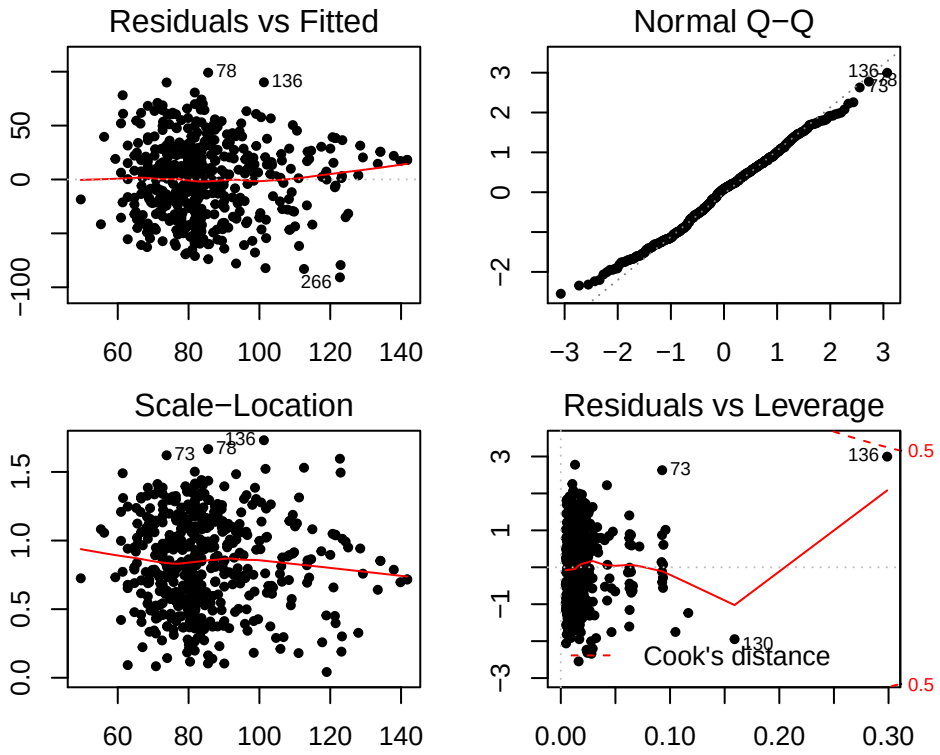
This plot suggest that people who are married tend to experience a rise in income between 1974 and 1978 while those are not married tend not to see this increase. This is likely a result of the fact that people who get married tend to be motivated to do so due to increased access to financial opportunities and therefore greater possibilities for income increase. As this difference in association between square rooted income in 1978 and income in 1974 is quite distint, it was determined that an interaction term between income in 1974 and the indicator for whether or not the person is married should be included in our model.

Linear Regression Modeling

Since the relationship between the earnings in 1978 (*re78_c*) and whether the participants received job training (*treat*) may depend on other covariates, the model with the interaction terms of *treat* and others is set to be the full model. To match the inferential purpose of this report while favoring a parsimonious model, stepwise model selection using the Bayesian information criterion (BIC) is applied to generate the final model.

After fitting the final model, however, the diagnostic plots indicate that a non-linearity seems to exist and the nomality assumption does not be validated due to the observation 97. Therefore, a square root transformation on the response variable is first performed based on result of the Box-Cox power transformation. Second, since the transformation does not resolve the normality issue, the observation 97 is excluded.





In the end, the final model is:

$$\hat{y}_{re78} = \hat{\beta}_0 \times treat + \hat{\beta}_1 \times black + \hat{\beta}_2 \times hispan + \hat{\beta}_3 \times educ_c + \hat{\beta}_4 \times re74_c + \hat{\beta}_5 \times treat \times black + \hat{\beta}_6 \times treat \times hispan + \hat{\beta}_7 \times treat \times re74$$

Table 1: Linear Regression Summary

Coefficients	Estimates	Std. Error	t-value	Pr(>|t|)	Signif
Intercept	8508.369	475.829	17.881	0.000	***
treat1 (received training)	379.169	1824.206	0.208	0.835	
black1	-710.167	1010.816	-0.703	0.483	
hispan1	1004.096	1104.257	0.909	0.364	
mean centered educ	361.027	127.336	2.835	0.005	**
mean centered real earning in 1974	0.449	0.055	8.147	0.000	***
received training : black1	609.748	2086.251	0.292	0.770	
received training : hispan1	-2256.707	2880.855	-0.783	0.434	
real earning in 1974 : received training	-0.333	0.124	-2.682	0.008	**

[1] Residual standard error: 36.1 on 459 degrees of freedom
[2] Adjusted R-squared: 0.165

Interpretation of the model coefficients

- The three most statistically significant predictors are: mean centered education, mean centered earnings in 1974 and the interaction term between training and earnings in 1974. The most significant variable on the basis of the t-statistic is mean centered earnings of 1974.
- Coefficient of re74_c: 0.0022. For every 10 dollar increase in mean centred real earnings of 1974, there is a 0.0223 unit increase in sqrt(dollar value of earning in 78), given that all other variables are 0 and education grade is centered at mean. This corresponds to a 4 ($ or unit) increase in the predicted dollar value of re78 (i.e. real earnings in 1978)
- Coefficient of educ_c: 1.89. For every 1 grade increase in mean centred education level, there is a 1.89 unit increase in sqrt(dollar value of earning in 78), given that all other variables are 0 and re74 earning is centered at mean. This corresponds to a 365 ($ or unit) increase in the predicted dollar value of re78 (i.e. real earnings in 1978).
- The adjusted R-squared is 16.1% which means 16.1% in the response (sqrt(re78)) was explained by the model. The standard error is 35.9 on 461 degrees of freedon which is high.
- The 95% confidence interval for mean-centered re74 is [0.0016,0.0028] for sqrt(re78). This means that the range of effect of the earnings in 1974 on the **square-root of earnings in 1978** is between [0.0016,0.0028]. Similarly, the range for education is [0.57,3.22].

Conclusion

1. Training is not a statistically significant factor for earnings in 1978. This implies there is no quantifibale impact of the training predictor on sqrt(re78) or re78.
2. However, it comes up to be statistically significant when observed in its interaction with the earnings in 1974. This would mean that people would have seen a change in their earnings in 1978 due to the training according to their earnings in 1974 before the training started. This is interesting because the effect is negative (on sqrt(re78)).
3. The demographic predictor that comes out to be statistically significant is the education variable. Education has a positive effect on sqrt(re78), and thus on re78 too.

Introduction

As a follow up to the analysis above, while the hypothesis that treatment impacts income in 1978 has been rejected, the question of whether or not treatment increases the odds of having a non zero wage in 1978 becomes of interest. Certainly, even if treatment does not affect the amount of wage increase, it may still be considered of value if it increases the odds of having a non-zero income in 1978 i.e. increases the odds of being employed versus unemployed. The following second section of this report considers the same subsection of the data from this study as before to explore additional question that the researchers who conducted the study may have considered:

- Is there evidence that workers who receive job training have greater odds of having a non-zero income in 1978?
- Can the impact of receiving job training on the odds of having a non-zero income be quantified? What is the likely range of the effect of the treatment?
- Does the effect of the treatment differ by race/demographics?
- What are other associations worth noting with the odds of having a positive income in 1978?

A logistic regression model with multiple predictors is developed to map the data and interpretted to posit answers to the afore mentioned questions of interest.

Data

Data Pre-Processing

The data used in this section of the analysis is the same as that used in the previous section with the following notable exceptions:

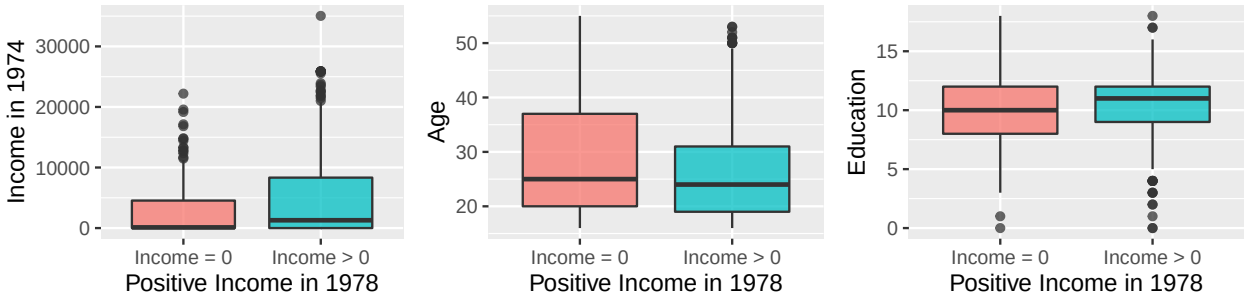
1. The data for income in 1978 was transformed into a binary outcome variable equalling 1 if the income was greater than zero (meaning that the person was employed) and 0 otherwise. The response variable (*re78*) as such was transformed into a binary variable.
2. Instead of using a subsection of the data (as before), in this portion of tha analysis, all the data was used.

The data for 1975 was yet again excluded from the analysis due to it representing an intermediate stage of this study and containing cases where participants were paid to join the study.

Exploratory Data Analysis

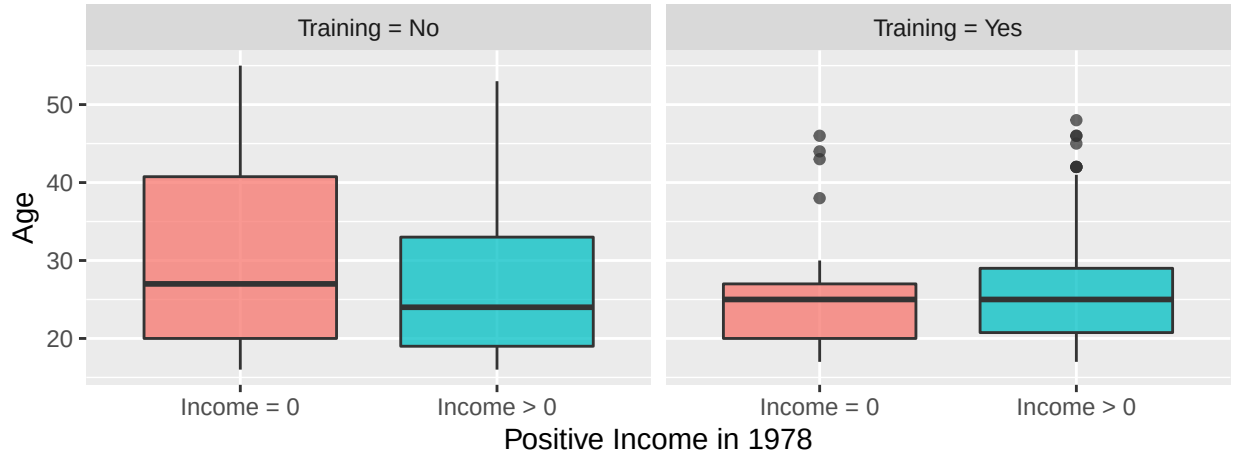
Before the model was developed, data exploration was undertaken to elucidate the relations and associations among the possible predictors. A large number of plots and summary statistics were generated to identify associations between the variables and make strong judgements about which predictors to include in the final model. Additionally, these plots and statistics were used to identify potential concerns within the data that may impact the analysis.

Firstly, boxplots of the continuous predictors, i.e. income in 1974, age, and education against the indicator for whether or not the particpant had a non-zero income in 1978 were produced to assess if the distributions of the data varried between categories, thereby suggesting an association between the predictor and the *re78* variable.



Among these variables, all three plots show some kind of association with *re78* and as such all three continuous variables were identified as potentially strong predictors of the odds of having positive income in 1978.

To detect the exitence of any possible interactions/associations among the predictor variables, frequency tables and additional box plots were produced and studied. In particular, the frequency tables for the association between treatment and *re74*, and this association segmented by each race were closely scrutinized for evidence of significance. However, evidence for their significance was not found during the exploratory data analysis process. On the other hand, it was discovered that the association between age and *re74* appears to vary by treatment:



This plot seems to suggest that the odds of having a non-zero income in 1978 are higher for older participants in the treatment group than those in the non-treatment group. As such, an interaction term between age and the indicator variable for treatment was considered as an additional predictor variable for the final model.

Logistic Regression Modeling

We first tried to fit a full logistic regression model by centering all the continuous predictors for better interpretation. The summary of the model indicates that only age, black and real annual earinings in 1974 appear to be significant. Among the three variables, age is very significant with a p value of less than 0.0003, while black and annual earnings in 1974 being moderate/slight significant and the rest not significant. When running the binned residual plots for the full model for the predicted probabilities and each variables, we do not see obvious pattern or more than 3 points falling out of the confidence interval that require special attention.

Based on the output of our full model, we decided to test the influence of variables including married, nodegree and education by dropping them one by one and running deviance tests to see if the model is being affected in anyway. Based

on the the result of our deviance tests, removing the three of them do not place a significant influence on the model. By looking at the binned residual plots again we do not observe too much of a difference from the full model.

When running model validation with confusion matrix and ROC curve with remained variables, we are getting an accuracy rate of *0.619* with sensitivity being *0.639*, specificity being *0.552* and AUC being *0.633*. To test interactions among different predictor variables, we made binned plots and ran different deviance tests to see if there are actually significant interactions. We started by including all interactions between paired variables in the model and removing each interaction one by one. In addition, by running a stepwise selection using BIC and setting treatment, the real annual earnings in 1974 and interactons between the treatment and demographic groups as the lowerbound baseline, the other predictor variables that the model selection retained are age, black, hispan, interaction between the treatment groups and age. This corresponds well to our previous analysis and EDA.

We present our final logistic model as the following:

$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_{treat1} treat1_i + \beta_{age} age_i + \beta_{black1} black1_i + \beta_{hispan1} hispan1_i + \beta_{re74} re74_i + \beta_{treat1:age} treat1_i : age_i + \beta_{treat1:black1} treat1_i : black1_i + \beta_{treat1:hispan1} treat1_i : hispan1_i$

Table 2: Logistic Regression Summary

Coefficients	Log-Odds Estimate	Odds Estimate	Log-Odds Std. Error	z-value	Pr(>|z|)	Signif
Intercept	1.339	3.815	0.156	8.594	0.000	***
Received Training	1.022	2.777	0.773	1.322	0.186	
Age	-0.053	0.949	0.011	-4.782	0.000	***
Black	0.000	1.000	0.000	3.712	0.000	
Hispanic	-0.393	0.675	0.293	-1.341	0.180	
Real Income 1974	0.068	1.071	0.364	0.187	0.852	***
Received Training : Age	0.075	1.078	0.027	2.759	0.006	**
Received Training : Black	-0.757	0.469	0.827	-0.915	0.360	
Received Training : Hispanic	14.406	1805712.491	716.789	0.020	0.984	

[1] Null Deviance = 666.50
[2] Residual Deviance: 623.15
[3] AIC: 641.15

Table 3: Logistic Regression Summary On The Odds Scale

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.82e+00	0.1558	8.5943	0.0000	2.8352	5.23e+00
treat1	2.78e+00	0.7726	1.3222	0.1861	0.7458	1.81e+01
age_c	9.49e-01	0.0111	-4.7815	0.0000	0.9278	9.69e-01
re74_c	1.00e+00	0.0000	3.7119	0.0002	1.0000	1.00e+00
black1	6.75e-01	0.2934	-1.3411	0.1799	0.3817	1.21e+00
hispan1	1.07e+00	0.3642	0.1870	0.8517	0.5377	2.27e+00
treat1:age_c	1.08e+00	0.0273	2.7594	0.0058	1.0235	1.14e+00
treat1:black1	4.69e-01	0.8272	-0.9151	0.3601	0.0671	2.00e+00
treat1:hispan1	1.81e+06	716.7894	0.0201	0.9840	0.0002	2.38e+127

Running again for the binned residual plots for our final model and it looks pretty reasonable. We are getting an accuracy rate of *0.617* with sensitivity being *0.618*, specificity being *0.615* and AUC being *0.652*. We also tested VIF on the model but did not detect any multicollinearity problems.

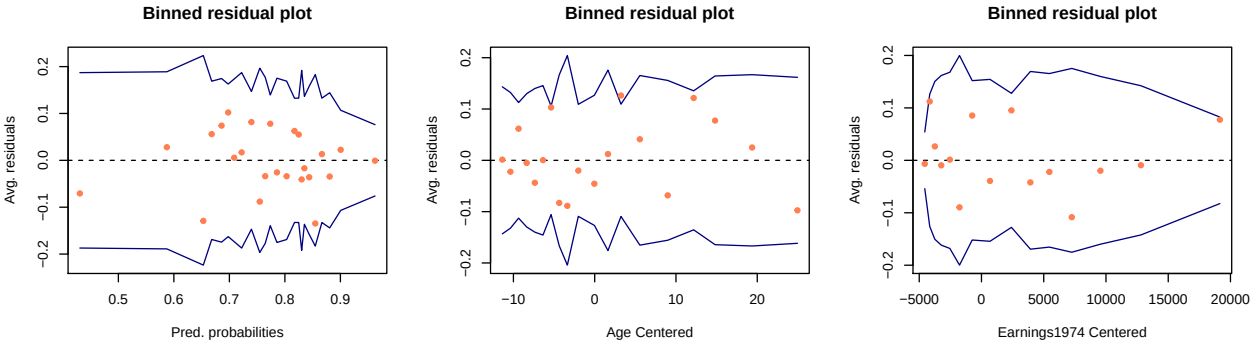
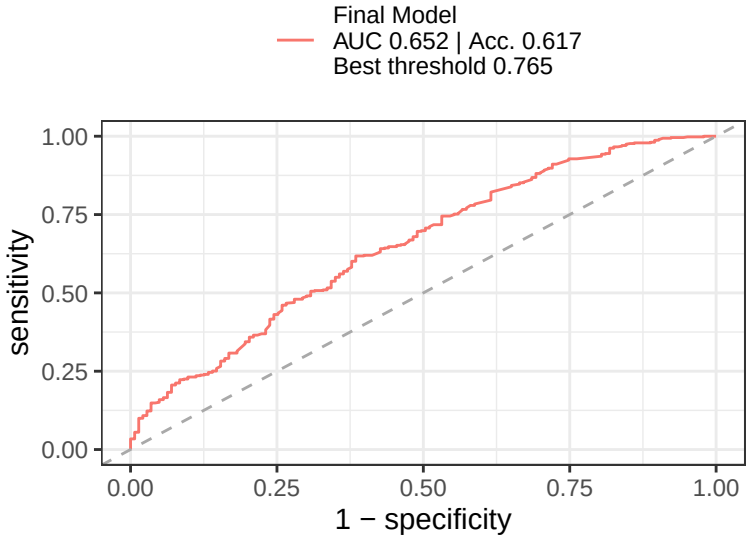


Table 4: Results of Confusion Matrix

Cutoff	0.765
Accuracy	0.617
Sensitivity	0.618
Specificity	0.615



Intepretation

- ***age*** has the highest influence on ***Positive Income*** based on z-value. The interaction between ***treat*** and ***age*** is also significant. Real income from 1974 is also very significant. The sensitivity of ***Positive Income*** to each of the significant factors (when everything else is held constant) is as follows:
  - ***Age*** - An increase in age of 1 year results in a 5% decrease in the odds of a positive income.
  - ***Real Income in 1974*** - $1 increase in real income in 1974 increases the odds of a Positive income in 1978 by 0.01%. The same applies for real income in 1975.
  - ***Training and Age*** - For a unit increase in age, those that received training will have their odds of receiving a Positive income in 1978 higher by 7.6%.
- Based on EDA, ANOVA and p-values above, there is no evidence that ***treat*** has a sigficant effect on ***Positive wages***. The demographic factors ***black*** and ***hispan*** as well as their interactions with ***treat*** are also not significant.

Conclusion

- Having considered the analysis presented above, there is no evidence of trained workers being more likely to have a positive wage relative to those who do not receive the training.
  - Hence it is not meaningful to quantify the effect of the training as a result.
- In addition, we could not establish any evidence that effects of training differed by demographic group.
- The investigation revealed potential differences in the effect of training over different ages. This will need to be investigated further.

Possible Limitations

1. The linear regression model possesses high standard errors and low R-squared value, which limit the predictive accuracy using the model.
2. The data is a non-random subset of the total population, using this biased sample may affect the inferential ability of the models.
3. Although the square-root transformation helps valiating the assumptions of linear models, it makes interpreting the coefficients and the fitted values more difficult.
4. The imbalanced data of the binary coded *re78* may harm its predictive accuracy.

References

[1] Akande, O. M. 2019. Team Project 1. Retrieved September 28, 2019 from https://akandelanre.github.io/IDS702_F19/project/team-project-01.html

Appendix

```
## ----include = FALSE-----------------------------------------------------
knitr::opts_chunk$set(echo=FALSE,
                      warning=FALSE,
                      message=FALSE,
                      fig.align="center",
                      fig.pos='H',
                      results="asis")


## ----results="hide"------------------------------------------------------
list_of_packages = c("tidyverse",
                     "dplyr",
                     "caret",
                     "broom",
                     "regclass",
                     "knitr",
                     "xtable",
                     "rlist",
                     "car",
                     "ggfortify",
                     "kableExtra",
                     "GGally",
                     "arm",
                     "pROC",
                     "gridExtra",
                     "forcats",
                     "magrittr",
                     "tibble",
                     "rms",
                     "MASS",
                     "stringr")
packages = list_of_packages[!(list_of_packages %in% installed.packages()[,"Package"])]
if(length(packages)){install.packages(packages)}
pkg_lib = lapply(list_of_packages,
                 require,
                 character.only=TRUE)
digit = 3
options(digits=digit,
        xtable.comment=FALSE)
```

```r
## ------------------------------------------------------------------------
data = read_csv("lalondedata.csv")
facs = c("treat", "black", "hispan", "married", "nodegree")
for (col in facs){
  data[[col]] = data[[col]] %>% as.character()
}
data[["treat"]][which(data[["re78"]] == 0)] = NA
#(data[["re78"]][which(data[["re78"]] == 0 & data[["treat"]] == 1)])/length(data[["re78"]][which(data[[
#length(data[["re78"]][which(data[["re78"]] == 0 & data[["treat"]] == 0)])/length(data[["re78"]][which(
data = data %>% na.omit()
data[["re78_root"]] = sqrt(data[["re78"]])


## ---- fig.height=2.2-----------------------------------------------------
p1 = data %>%
      ggplot(aes(x = re78)) +
      geom_density(fill = "blue", alpha = 0.5) +
      xlab("Income in 1978") +
      ggtitle("Density Plot") +
      theme(plot.title = element_text(hjust = 0.5))
p2 = data %>%
      ggplot(aes(x = log(re78))) +
      geom_density(fill = "red", alpha = 0.5) +
      xlab("Log(Income) in 1978") +
      ggtitle("Density Plot") +
      theme(plot.title = element_text(hjust = 0.5))
p3 = data %>%
      ggplot(aes(x = sqrt(re78))) +
      geom_density(fill = "green", alpha = 0.5) +
      xlab("Sqrt(Income) in 1978") +
      ggtitle("Density Plot") +
      theme(plot.title = element_text(hjust = 0.5))
grid.arrange(p1, p2, p3, ncol = 3)


## ----fig.height=2.2------------------------------------------------------
data[["black"]][which(data[["black"]] == 0)] = "Black = No"
data[["black"]][which(data[["black"]] == 1)] = "Black = Yes"
data[["hispan"]][which(data[["hispan"]] == 0)] = "Hispanic = No"
data[["hispan"]][which(data[["hispan"]] == 1)] = "Hispanic = Yes"
data[["married"]][which(data[["married"]] == 0)] = "Married = No"
data[["married"]][which(data[["married"]] == 1)] = "Married = Yes"
data[["treat"]][which(data[["treat"]] == 0)] = "No"
data[["treat"]][which(data[["treat"]] == 1)] = "Yes"

p1 = data %>%
```

```
  ggplot(aes(x = treat, y = re78_root, fill = treat)) +
  geom_boxplot(alpha = 0.75) + facet_wrap(~black) +
  ylab("Sqrt(Income) in 1978") +
  xlab("Job Training Provided") +
  ggtitle("Sqrt(Income) vs Treatment") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

p2 = data %>%
  ggplot(aes(x = treat, y = re78_root, fill = treat)) +
  geom_boxplot(alpha = 0.75) + facet_wrap(~hispan) +
  ylab("Sqrt(Income) in 1978") +
  xlab("Job Training Provided") +
  ggtitle("Sqrt(Income) vs Treatment") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

grid.arrange(p1, p2, ncol = 2)


## ----fig.height=2.2-------------------------------------------------------
data %>%
  ggplot(aes(x = re74, y = re78_root)) +
  geom_point() + geom_smooth(method = "lm") +
  facet_wrap(~married)  +
  ylab("Sqrt(Income) in 1978") +
  xlab("Income in 1974") +
  ggtitle("Sqrt(Income) in 1978 vs Income in 1974") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")


## ----fig.height=2.2-------------------------------------------------------
order_cols = function(df, col.order){
  df %>%
    dplyr::select(sapply(., class) %>%
                    .[order(match(., col.order))]
                  %>% names)}
centering = function(x){
  x = x - mean(x)
  return(x)
}

lalonde = read_csv("lalondedata.txt")
# sum(ifelse(lalonde$re78==0, 1, 0))
rm.var = c('X1', 're75')
y = lalonde %>%
    dplyr::select('re78') %>%
    mutate_at(vars(starts_with('re78')), list( ~ na_if(., 0))) %>%
    na.omit()
```

```
lalonde = lalonde %>%
    mutate_at(vars(treat,
                   black,
                   hispan,
                   married,
                   nodegree),
              as.factor) %>%
  dplyr::select(-rm.var) %>%
  mutate_at(vars(starts_with('re78')), list( ~ na_if(., 0))) %>%
  na.omit() %>%
  mutate_if(is.numeric , centering) %>%
  order_cols('factor') %>%
  dplyr::select(-re78) %>%
  bind_cols(y) %>%
  rename(re74_c = re74,
         educ_c = educ)


## ----echo=FALSE, results='hide'-----------------------------------------
# generating a full model: main effects + interaction terms of main effects
full.model = lm(re78 ~ . + treat:., lalonde)
n = nrow(lalonde)
# BIC
final.model = step(full.model,
                   scope=list(lower = ~ re74_c + treat +
                                treat*black + treat*hispan),
                   k=log(n),
                   direction='both')


## -----------------------------------------------------------------------
# par(mfrow=c(2,2),
#     mar=rep(2,4))
# plot(final.model, pch=20)


## ----fig.height=3.5, fig.width=4----------------------------------------
b = MASS::boxcox(final.model)
# bc = cbind(b$x, b$y)
# sorted_bc = bc[order(-b$y),]
# lambda = sorted_bc[1]


## ----fig.height=4, fig.width=5------------------------------------------
final.fit = lm(sqrt(re78) ~ treat + black + hispan +
                   educ_c + re74_c + treat:black +
```

```
                  treat:hispan + treat:re74_c,
    data = lalonde[-97,])
par(mfrow=c(2,2),
    mar=rep(2,4))
plot(final.fit, pch=20)
# summary(final.model)


## ------------------------------------------------------------------------
# final.model %>%
#     tidy(conf.int=T, digit=3) %>%
#     kable(format="latex", caption="Linear Regression Summary", booktabs = TRUE)


## ------------------------------------------------------------------------
sm = summary(final.model)[["coefficients"]] %>% as_tibble() %>% round(4)
Coefficients = c("Intercept", "treat1 (received training)", "black1", "hispan1",  "mean centered educ",
sm = sm %>% add_column(Coefficients, .before = "Estimate")

sm[["Significance"]] = c(rep("***", 1), rep("", 3),  rep("**", 1), rep("***", 1), rep("", 2),rep("**",
sm %>%
  knitr::kable("latex", booktabs = T, linesep = "",
        escape = F, caption = "Linear Regression Summary",
        align = c('l', rep('r', 5), 'c'),
        col.names = linebreak(c(
          "Coefficients",
          "Estimates",
          "Std. Error",
          "t-value",
          "Pr(>|t|)",
          "Signif"
        ))) %>%
  kable_styling(full_width = F,latex_options = "hold_position") %>%
  footnote(
    number = c("Residual standard error: 36.1 on 459 degrees of freedom",
               "Adjusted R-squared:  0.165"),
    general_title = " "
  )


## ----fig.height=2.2------------------------------------------------------
data = read_csv("lalondedata.csv")
facs = c("treat", "black", "hispan", "married", "nodegree")
for (col in facs){
  data[[col]] = data[[col]] %>% as.character()
}
data[["re78"]][which(data[["re78"]] > 0)] = "Income > 0"
```

```
data[["re78"]][which(data[["re78"]] == 0)] = "Income = 0"
data[["black"]][which(data[["black"]] == 0)] = "Black = No"
data[["black"]][which(data[["black"]] == 1)] = "Black = Yes"
data[["hispan"]][which(data[["hispan"]] == 0)] = "Hispanic = No"
data[["hispan"]][which(data[["hispan"]] == 1)] = "Hispanic = Yes"
data[["married"]][which(data[["married"]] == 0)] = "Married = No"
data[["married"]][which(data[["married"]] == 1)] = "Married = Yes"
data[["treat"]][which(data[["treat"]] == 0)] = "Training = No"
data[["treat"]][which(data[["treat"]] == 1)] = "Training = Yes"


## ----fig.height=2, fig.width=8-------------------------------------------
p1 = data %>%
  ggplot(aes(x = re78, y = re74, fill = re78)) +
  geom_boxplot(alpha = 0.75) +
  xlab("Positive Income in 1978") +
  ylab("Income in 1974") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

p2 = data %>%
  ggplot(aes(x = re78, y = age, fill = re78)) +
  geom_boxplot(alpha = 0.75) +
  xlab("Positive Income in 1978") +
  ylab("Age") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

p3 = data %>%
  ggplot(aes(x = re78, y = educ, fill = re78)) +
  geom_boxplot(alpha = 0.75) +
  xlab("Positive Income in 1978") +
  ylab("Education") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

grid.arrange(p1, p2, p3, ncol = 3)


## ----fig.height=2.5------------------------------------------------------
data %>%
  ggplot(aes(x = re78, y = age, fill = re78)) +
  geom_boxplot(alpha = 0.75) +
  xlab("Positive Income in 1978") +
  ylab("Age") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none") + facet_wrap(~treat)


## ------------------------------------------------------------------------
rm(list = ls())
```

```
###### Load the data
lalonde <- read.csv("lalondedata.txt",header=T)
lalonde$incomebool <- ifelse(lalonde$re78>0, 1, 0)
lalonde$treat <- as.factor(lalonde$treat)
lalonde$black <- as.factor(lalonde$black)
lalonde$hispan <- as.factor(lalonde$hispan)
lalonde$married <- as.factor(lalonde$married)
lalonde$nodegree <- as.factor(lalonde$nodegree)


## ----message=FALSE, echo=FALSE, warning=FALSE, results='hide'------------
lalonde$age_c = lalonde$age - mean(lalonde$age)
lalonde$re74_c = lalonde$re74 - mean(lalonde$re74)
lalondereg6 = glm(incomebool ~ (treat + age_c + re74_c + black + hispan + educ)^2 , data = lalonde, fami
summary(lalondereg6)
n = nrow(lalonde)
final.model = step(lalondereg6, direction ="both", scope=list( lower = ~ re74_c + treat*hispan + treat*b
summary(final.model)


## ----echo=FALSE----------------------------------------------------------
odds = final.model$coefficients %>% exp() %>% round(4) %>%
  format(scientific=F) %>% str_trim()
sm = summary(final.model)[["coefficients"]] %>% as_tibble() %>% round(4)
Coefficients = c("Intercept", "Received Training", "Age", "Black", "Hispanic", "Real Income 1974", "Rec
sm = sm %>% add_column(Coefficients, .before = "Estimate") %>%
  add_column(odds, .before = "Std. Error")

sm[["Significance"]] = c(rep("***", 1), rep("", 1), rep("***", 1), rep("", 2),rep("***",1),"**",rep("",1
sm %>%
  kable("latex", booktabs = T, linesep = "",
        escape = F, caption = "Logistic Regression Summary",
        align = c('l', rep('r', 5), 'c'),
        col.names = linebreak(c(
          "Coefficients",
          "Log-Odds\nEstimate",
          "Odds Estimate",
          "Log-Odds\nStd. Error",
          "z-value",
          "Pr(>|z|)",
          "Signif"
        ))) %>%
  kable_styling(full_width = F, latex_options = "hold_position") %>%
  footnote(
    number = c("Null Deviance = 666.50",
               "Residual Deviance: 623.15",
               "AIC: 641.15"),
```

```
    general_title = "  "
  )


## ----message=FALSE, echo=FALSE, warning=FALSE-----------------------------
kable(tidy(final.model,conf.int = TRUE, exponentiate=TRUE),
      "latex", booktabs = T, digits = 4,
      caption="Logistic Regression Summary On The Odds Scale") %>%
  kable_styling(latex_options = "hold_position")


## ----echo=FALSE, fig.width=10,fig.height=3--------------------------------
rawresid1 <- residuals(final.model,"resp")
par(mfrow=c(1,3))
binnedplot(x=fitted(final.model),y=rawresid1,xlab="Pred. probabilities",
           col.int="navy",ylab="Avg. residuals",main="Binned residual plot",col.pts="coral")

binnedplot(x=lalonde$age_c,y=rawresid1,xlab="Age Centered",
           col.int="navy",ylab="Avg. residuals",main="Binned residual plot",col.pts="coral")

binnedplot(x=lalonde$re74_c,y=rawresid1,xlab="Earnings1974 Centered",
           col.int="navy",ylab="Avg. residuals",main="Binned residual plot",col.pts="coral")


## ----echo=FALSE,message=FALSE, fig.height=3, fig.width=4-----------------
prob = predict(final.model, type="response")
roc.curve = roc(lalonde$incomebool ~ prob, lalonde)
cutoff.best = coords(roc.curve, "best",
                     ret=c("threshold", "specificity", "sensitivity"),
                     transpose = TRUE)
pred = ifelse(prob>cutoff.best[1], 1, 0)
acc = mean(pred==lalonde$incomebool) %>% round(3)
str = paste(paste("Final Model\n", "AUC", sep=""),
             round(roc.curve$auc, 3), "|", "Acc.", acc,
            "\nBest threshold", paste(cutoff.best[1] %>% round(3)))
ggroc(list(roc.curve),
      legacy.axes=TRUE) +
  theme_bw() +
  labs(colour='') +
  theme(legend.position="top") +
  geom_abline(intercept = 0, slope = 1,
               color = "darkgrey", linetype = "dashed") +
  scale_color_manual(name='',
                     labels=c(str),
                     values=c("#F8766D"))
```

```
## ----code = readLines(knitr::purl("Team Project 1.Rmd", documentation = 1)), echo = T, eval = F----
## NA
```