Team Project 1

Pranav Manjunath (Checker, Coordinator) Aiman Haider (Presenter)
Xinyi Pan (Programmer) Maobin Guo (Writer)

Summary

This report analyzes the impact of a training program by examining its effect on the real annual earnings and its likeliness to earn a non-zero wage. To do so, we use linear and logistic regression models, respectively. We use a dataset from the National Supported Work (NSW) Demonstration (1975-1978)[1][2] for the same and build models using the AIC Forward criterion based on the statistical significance and reasonableness. We then find out that the variables treat(taking training), age, marital status, education, age given marital status, and age given training have significant associations with annual incomes, suggesting that taking the training might be a very important factor associated with an increase in wages. Also, we find that non-zero wages are associated with race, age, marital status, race given marital status, age given nodegree and age given training suggesting that training is not a significant factor associated with non-zero wages by itself. Thus, it can be understood that the training program as a factor does seem to have a very strong association with an increase in wages. However, it is also found that it does not have much association with non-zero wages and is associated with other demographic factors and educational qualifications.

Introduction

In the 1970s, researchers in the United States ran several randomized experiments to evaluate public policy programs. One of the most famous experiments was the National Supported Work (NSW) Demonstration. Researchers wanted to assess whether or not job training for disadvantaged workers affected their wages. Based on a subset of this dataset, to understand the impact of the training program, we need to look at two main questions: Part I: Is there evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training? To address this question, we need to quantify the effect of the treatment: receiving job training on real annual earnings and understanding the likely range for the effect of training. We also need to check if there is any evidence that the effects differ by demographic groups and other interesting factors.

Part II: Is there evidence that workers who receive job training tend to be more likely to have positive (non-zero) wages than workers who do not receive job training? To understand this question, we need to quantify the effect of the treatment, that is, receiving job training, on the odds of having non-zero wages, and what would be the likely range for the effect of training. We also need to see if there is any evidence that the effects differ by demographic groups and also if there are other interesting associations with positive wages. These questions would help us understand the potential association of the impact of the training program on wages. To answer these questions, the report uses a linear regression on the differences in wages and a logistic regression model on the odds of getting a non-zero wage after the training. It begins with an EDA of the data. It tries to build a model by exploring models built with the help of AIC criteria using forwarding model building and choosing the most suitable model based on accuracy and plausibility to answer the above questions.

[1] Evaluating the Econometric Evaluations of Training Programs with Experimental Data

[2] Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs

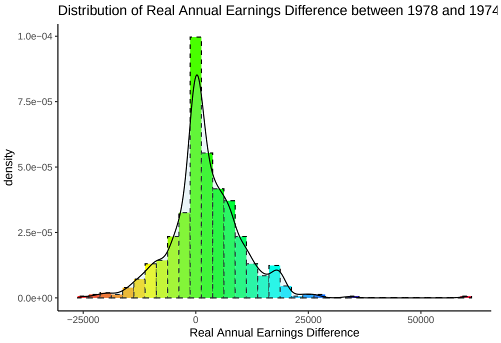
Part I

DATA

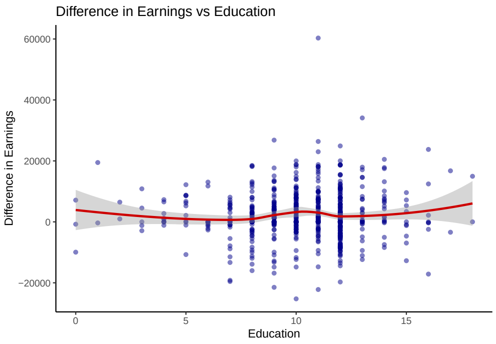
The data we have considered for the analysis consists of 614 observations and 11 variables viz. X (Participant ID), treat, age, educ (education level), black, hispan (Hispan), married, nodegree, re74 (real earnings in 1974), re75 (real earnings in 1975), and re78 (real earnings in 1978). Further, the dataset contains 185 male participants who attended the training program and 429 male participants who have not attended the training program. There are 243 black and 72 hispanic participants. We converted treat, black, hispan, married, nodegree into binary factor variables, educ to a discrete variable and age as a continuous variable. The columns X and re75 will not be used in this analysis. We will not use re75 as it cannot be considered as a constant baseline among the participants.

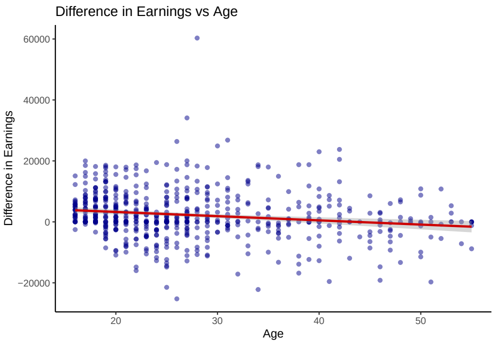
Exploratory Data Analysis

Firstly, we look at the response variable. We notice that re78 is not normally distributed so we instead choose the difference between re78 and re74 for two reasons: (i) This seems to mimic normal distribution more closely than the former and (ii) It is the variable of direct interest in the analysis.

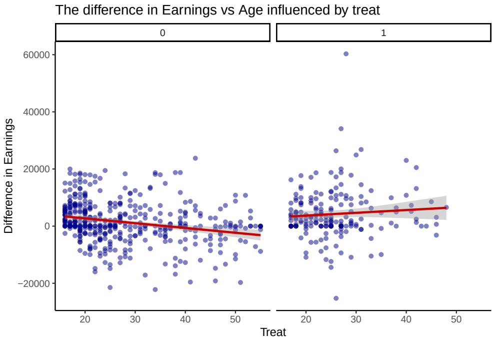


Further, It is observed that the variable age shows a slightly decreasing linear relationship with the wage differences. While education shows a non-linear trend. We also see that categorical variables treat, black, hispan and married seem to show some differences based on the categories. We then look into the interactions and find that the association between wage differences and treat seems to differ by age; It shows a decreasing trend for non-trainees while an increasing trend for trainees. The trend between the two also shows differences by races, educ and married. We also find that trends between the wage differences and education differ by married. This is also found so for the age variable to differ by married.





Unlike the relationship between age and treat (Right), the relationship between education(left) and treat is not linear. It indicates that some non-linear transformation might be confirmed, which is later found to be the case with residuals with only linear terms.



The plot above illustrates the interaction between age and treat versus the response variable. The annual earnings trend with the increase of age is different according to treat. Further investigation of this interaction would be exerted in the model fitting step.

Model

Model selection

Based on the above discussion, we try building a model using the AIC forward and selecting the model.

Null Model contains only the treat variable while the full model contains all the predictors along with all the possible interactions among predictors. The null model has an AIC value of 12771.89 while the full model has an AIC value of 12764.68. The output of Forward AIC model is:

$$\begin{aligned} diff_i = &\beta_0 + \beta_1 * treat + \beta_2 * agec + \beta_3 * married + \beta_4 * educc + \\ &\beta_5 * educc2 + \beta_6 * treat : agec + \beta_7 * agec : married + \varepsilon_i ; \varepsilon \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \end{aligned}$$

where agec stands for age centered variable and educc stands for education centered variable.

The AIC model contains the variables, treat, agec, married, educc, and educc2 (squared term of educ) along with the interaction between treat & agec and agec & married.

On comparing the model to the Null Model and Full Model, we find that the above model is more statistically significant (p-value: 6.668e-07 (null) and p-value:0.3817 (full)). Further, we test excluding and including some variables and found this to indeed be the most statistically significant one. However, we add the variables black and hispan as is required for analyzing the question of interest. We also add the variable educc as we retain its second order term. Interaction between “married” and “age” discovered during this process was also tested for by the ANOVA. Both its p-value and statistical significance through ANOVA indicate that it must be retained in the model.

Intuitively too, one can understand that wages depend on age and education. For age, it can be inferred that those who are elder tend to make lesser wages when compared to the youth in specific job fields. Further, interactions between age and training and age and marriage also make sense as they affect different sets of employment options e.g. a young bachelor can take more risk for wages.

Also of the two variables defining educational qualification- education and nodegree,we retain education duration was since it has a higher p-value compared to no-degree.

Final model

$$\begin{aligned} diff_i = &\beta_0 + \beta_1 * black_i + \beta_2 * hispan_i + \beta_3 * agec_i + \beta_4 * married_i + \beta_5 * treat_i : agec_i + \\ &\beta_6 * educc_i + \beta_7 * educc_i^2 + \beta_8 agec : married_i + \varepsilon_i ; \varepsilon \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \end{aligned}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1743.6676	655.9077	2.66	0.0081
treat1	3254.1008	884.7055	3.68	0.0003
black1	-698.6038	853.5575	-0.82	0.4134
hispan1	623.2948	1047.1434	0.60	0.5519
agec	-220.0966	52.3925	-4.20	0.0000
married1	-1879.6013	727.0043	-2.59	0.0100
educc	151.2465	131.5212	1.15	0.2506
educc2	55.4840	26.6443	2.08	0.0377
treat1:agec	300.3163	89.6793	3.35	0.0009
agec:married1	137.9935	71.1252	1.94	0.0528

Table 1: Coefficient-Level Estimates

	pValue	RSquare
value	7.25e-09	0.09

Table 2: Evaluationl

Model Summary

The baseline values incorporated in the intercept are treat = 0,black = 0,hispan = 0,age = 27, and married = 0. Keeping other things constant, a 27-year-old bachelor who was neither black or hispanic and received education for about 10 years is expected to earn $1,743 if he didn't receive the job training.

From the above model summary, we find that treat, agec, married, educc2 (square term of education) and the interactions of age and treat are strongly associated (statistically significant) with the differences in wages at the 0.05 level. The interaction between age and married is significant at the 0.1 significance level. With demographic factor, we notice that race (black or hispanic) is not significant with the difference in wages, p value greater than 0.1.

1. Controlling other factors, if the participant takes part in the training program, his real earning is likely to increase by $3254.1 over 4 years on average (p<0.001). The likely range of this value is ($1516.63, $4991.57). The average annual increase would be $813.5.

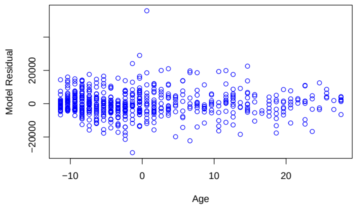
2. Controlling other factors, a unit increase of age from the mean value (27 years) tend to decrease the real earnings by $220.1 over 4 years on an average (p<0.0001). The likely range of it is (-$322.99, -$117.20). The average annual decrease would be $55.

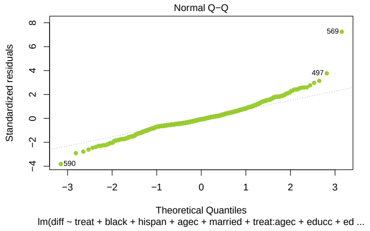
3. Controlling other factors, if the participant is married, his earnings tend to decrease by $1879.6 (p=0.01) over 4 years. The likely range of it is (-$3307.36, -$451.84). The average annual decrease would be $469.9.

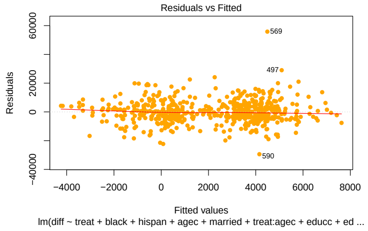
4. At the significance level of p<0.001, keeping other variables constant for every increase in age of a trainee (took the training program) his income tends to increases by $75 annually over and above the main effects.

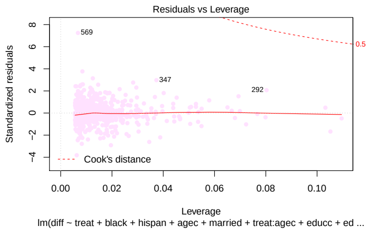
5. At the significance level of p<0.1, keeping other variables constant for every increase in age of a married worker, his income tends to increases by $34.5 annually over and above the main effects.

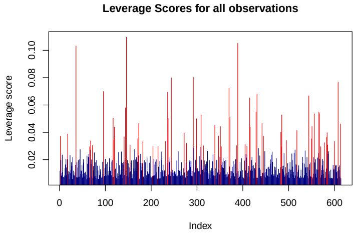
Model Assessment











From the residual plots, there are no violations of assumptions:

1) Linearity: The residual versus predictor plot seems random for the variable "age" and "educc2", while the other predictors are categorical.
2) Independence and Equal Variance: Absence of any pattern, randomness and wide-spread distributions over the spectrum support these assumptions.
3) QQ-plot supports the assumption of Normality generally as the plot is a straight line.

Outliers and High Leverage There are a few outliers and some high leverage points in model, however according to cook's distance they are not high influence points. And the model does not improve even after removing the outliers and hence we have not removed these points from the model.

Multicollinearity The variance inflation factor is less than 5 for all of the predictors. Therefore, it is safe to conclude that our final model is free from multicollinearity and can be used for our analysis.

Discussions

1. Is there evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training?

- From the linear regression model, the variable treat is statistically significant with a positive value slope concluding that there is evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training.

2. Quantify the effect of the treatment, that is, receiving job training, on real annual earnings.

- Controlling other factors, if the participant takes part in the training program, his real earning is likely to increase by $3254.1 over 4 years on average ($p<0.001$) at the 95% confidence interval.

3. What is a likely range for the effect of training?

- Keeping other variables constant, the effect of training on a participant may increase the real earning by a minimum of $1516.63 and a maximum of $4991.57 for 4 years.

4. Is there any evidence that the effects differ by demographic groups?

- We are considering age and race to be considered as demographic factors. Using the linear regression model, race (black and hispan) does not seem to be statistically significant with the response variable. Hence there is no evidence that the difference of real earnings from 1978 and 1974 differ by race.

However, age seems to be statistically significant (low p value) and has a negative slope. This indicates that for a unit increase in age, the real earnings over 4 years would tend to decrease by $220.10. The interaction of treat and age is significant in our model. Workers who received training would receive $124 per year for per 1-year increase in age, while the no-training workers' salary would decrease by $322 per year for per 1-year increase in age.

5. Are there other interesting associations with wages that are worth mentioning?

- Along with Treat and the demographic factors, the participant's marital status, education squared term, and the interaction between treat & agec and agec & married seems to be statistically significant. The worker's martial status has a negative slope indicating that a married participant tends to have a decrease in wages over four years. Education duration would increase workers' salaries. Given the quadratic effect of the education variable, the worker's salary is expected to increase faster as he received more than nine years of education. Unfortunately, he will be adversely affected if his education background is less than 9 years, as shown in the comparison plot. There is interaction effects between the worker's age and marital status on his earnings. With all other factors controlled, the job applicant's married status can offset the age discrimination and provide him a higher estimated salary.

Conclusion

To understand the impact of job training on the difference in wages from 1978 to 1974, we built a multiple linear regression model. We confirmed the model using AIC forward selection and ANOVA tests. The linear regression assumptions for the model were checked and we tested for potential outliers. To ensure no multicollinearity, VIF scores were generated, noticing that all variables had VIF value below 5. We noticed that the variable treat is statistically significant, inferring that job training positively influences the increase in real earnings from 1974 to 1978. In terms of the effect of demographic factors on the response variable, race does not seem to be statistically significant while age seems to be. Interestingly, marital status and education have strong associations with the difference in wages. Two interactions, Age & treat, age & married are also statistically significant as determined by p value less than 0.05. The effect of age is hard to interpret because it involves square transformation and interaction with treat.

Limitations: The final model's R-squared is only 0.088, which is relatively low, 8.8% proportion of variation in the response variable is being explained by the regression model.

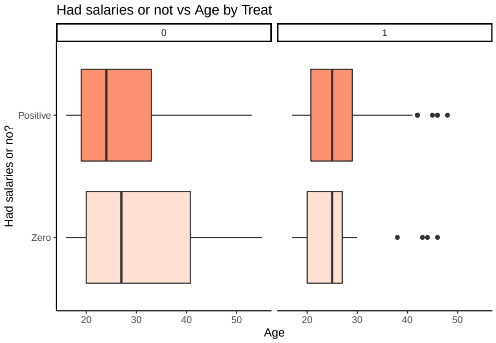
Part II

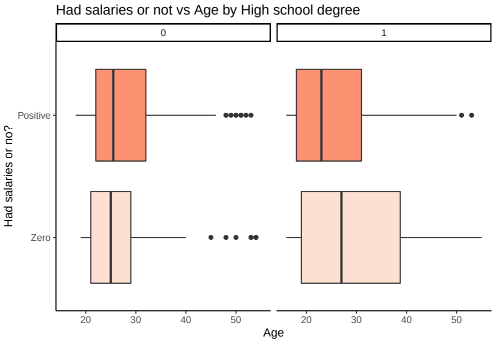
DATA

To answer the questions of interest, we converted the re78 variable into a binary response variable. A 0 value of the variable was taken to indicate that the participant made no earnings in 1978 while a 1 indicated some (non-zero) monetary earnings in 1978. In other words, the variable is an indicator of the participants' employment status in 1978 and is made our response variable. The variable re74 (real earnings in 1974) has also been removed from the analysis as the main aim was to look at the impact of training on positive wages (employment status). The other variables have been retained, as was also mentioned in Part 1.

Exploratory Data Analysis

We then look at some of the trends between the other variables and the response variable. On some analysis, we find that this employment variable seems to show relation with age.





As reported by the first plot (left), the relation between age and employment in 1978 is different based on whether or not the employee took the training program. The second plot points out that the relation between employment and age is also influenced by whether or not the worker has a high school degree

	0	1
Zero	0.26	0.21
Positive	0.74	0.79

Table 3: Employed rate with marital status - Non Hispanic

	0	1
Zero	0.08	0.28
Positive	0.92	0.72

Table 4: Employed rate with marital status - Hispanic

We also observe through the two tables shown that there are some differences in the relationship between marital status and positive wages according to race. For the Hispanic participants, the unmarried workers' unemployment rate seems to be lower than single workers of other races (8% vs. 26%). At the same time, this advantage disappears in the married Hispanic workers. The married Hispanic workers' unemployment rate is 28%, while Non-Hispanic races' married unemployment rate is 21%. This interesting sign of interaction would be further investigated in the model fitting.[3]

[3] The Chi-square tests for independence are not significant. The variables are explored later.

Model

Model selection

Based on the above trends, we build a logistic regression model using the AIC forward model selection method. For doing so we construct our Null Model as:

$$logit(\pi_i) = \beta_0 + \beta_1 * treat_i + \varepsilon_i$$

The full model includes all the variables and possible interactions among them.

Using the AIC forward model, we find that the square transform of "age" is significant in the model. After verification with ANOVA test, we decide to keep this transformation in the final model

As shown in the model output, the interactions between "age" &"nodegree" and "hispan" & "married" seems to be statistically significant, with a p value of 0.03 and 0.01 respectively. Interestingly, even though the interaction between treat and age is weakly significant (p-value=0.09), the ANOVA test confirmed its significance and hence we have included this in the final model. Although the variable treat is statistically not significant (p-value=0.4), we preserve it in our final model as it is used to answer the questions of interest.

Final model

$$\begin{aligned} logit(\pi_i) = &\beta_0 + \beta_1 * treat_i + \beta_2 * black_i + \beta_3 * agec_i + \beta_4 * agec_i^2 + \beta_5 * nodegree_i + \beta_6 * hispan_i + \\ &\beta_7 * married_i + \beta_8 * agec_i : nodegree_i + \beta_9 * hispan_i : married_i + \beta_{10} * agec_i : treat_i + \varepsilon_i \end{aligned}$$

Model Summary

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.4481	0.2641	5.48	0.0000
treat1	0.2360	0.2845	0.83	0.4069
black1	-0.5399	0.2648	-2.04	0.0415
agec	0.0138	0.0245	0.56	0.5734
agec2	-0.0021	0.0011	-1.93	0.0536
nodegree1	-0.0157	0.2147	-0.07	0.9417
hispan1	1.1998	0.6436	1.86	0.0623
married1	0.3621	0.2629	1.38	0.1683
agec:nodegree1	-0.0447	0.0210	-2.13	0.0330
hispan1:married1	-1.8004	0.7703	-2.34	0.0194
treat1:agec	0.0456	0.0276	1.65	0.0990

Table 5: Coefficients

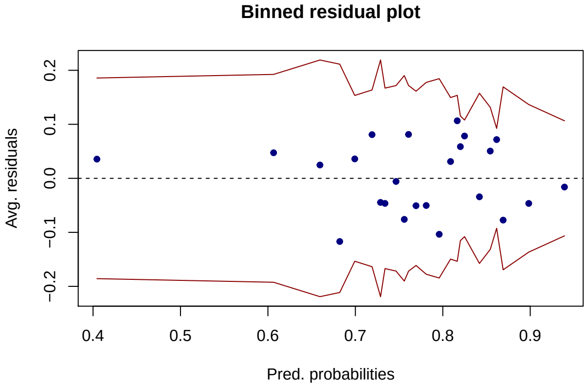
	NULL_deviance	Residual_deviance
1	666.50	627.54

Table 6: Deviance

The baseline values taken in the intercept is treat=0 and black=0, nodegree=0, hispan=0, age=27, and married=0. From the above, it can be seen that the variables black, agec2, hispan1, treat:agec, agec:nodegree and hispan:married. At the significance level 0.1, it can be observed that if a person is black then his odds of getting a positive wage decreases by 42% while when the person is a hispanic it increases by 232% compared to the baseline of being a non-black non-hispanic male. Further, for every unit increase in age if the person undertook the training then the odds increases by 4.6% and for every unit increase in age if the person does not have a degree then the odds decreases by 4.3%, keeping other things constant.

Model Verification

Residuals



The binned plot above illustrates the average residuals vs predicted probabilities. As shown, more than 95% of points reside inside the red band (within 95% confidence intervals) and no distinct pattern is observed. The binned plots for average residuals vs individual variables have been plotted and these plots also do not violate the above rules. This concludes a strong justification for the model's efficiency.

Null model deviance : 666.5. Final model deviance : 627.54. The deviance of our final model decreased by 38.96, best among the models we examined.

Multicollinearity: According to the table, most of the VIF values for the predictors are below 5, indicating moderately correlated. However, age centered seems to have a VIF value of 6.8, indicating high correlation. However, given that VIF of all the predictors are less than 10, it is safe to conclude that there is no significant multicollinearity among variables in the model.

To summarize, our final model can be used to answer our questions of interest in light of the above discussion.

Model Assessment

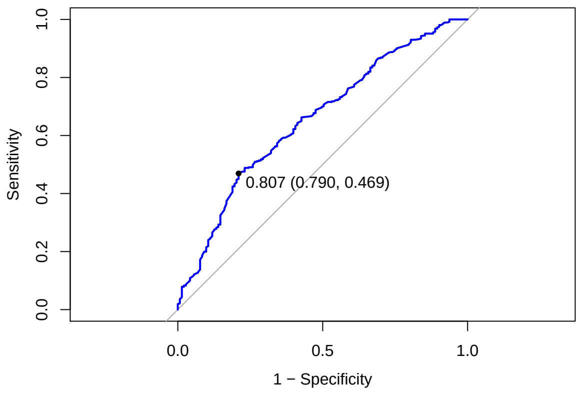
	0	1
0	91	192
1	52	279

Table 7: Confusion Matrix

The Sensitivity value, also known as true positive rate $\frac{TP}{TP+FN}$ of the model is 0.59.

The Specificity value, also known as true negative rate $\frac{TN}{FP+TN}$ of the model is 0.64

The Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$ of the model is 60%.



The ROC curve above has a AUC value of 0.65 and a cut-off value of 0.81, which was the best among the models we tested.

Discussion

1. According to the model, the variable treat is not statistically significant as it has a p value greater than 0.05. Hence receiving job training would not directly influence the odds of these worker's positive wages and cannot quantify the effect of treatment on the odds of having non-zero wages. However, the interaction item "treat1:agec" is relatively significant. At the significance level of 0.1, keeping other variables constant for every increase in age of a trainee the odds of getting a job (receiving earnings in 1978) increases by 4.67% income and above the main effects.
2. There is evidence that the effects differ by demographic groups. In the logistic regression model, black and hispan are statistically significant. For a black worker, his odds of getting a job are 41% lower than a non-black worker under the same other conditions. The odds ratio of a hispanic employee having positive earnings in 1978 increases by 2.32 times when compared to a non-hispanic employee.
3. Hispanic Race and marital status. Controlling other factors, a Hispanic workers odd to be employed is 231% higher than others. However, the effect would be brought down by being married. A married Hispanic worker's odds of getting a job are 83.4% lower.
4. Age is a complex factor in our model. Generally speaking, the coefficient of age seems to accelerate the trend of the workers' odds of not havng a job. Taking the job training may slow down the decreasing trends.

Conclusion

To understand the impact of job training on positive wages, we built a logistic regression model. We confirmed the model using AIC forward selection and ANOVA tests. The assessment of the model was done using binned residual plots and identifying specificity, sensitivity, accuracy, and area under the ROC. We noticed that the variable treat is not statistically significant, inferring that job training would not directly influence the odds of these worker's positive wages. Race and Age seemed to be statistically significant when predicting the positive wages. Three interactions, Age & treat, hispan & married, age & "no degree" are also statistically significant as determined by p value less than 0.05. The effect of age is hard to interpret because it involves square transformation and interaction with treat.

Limitation: Sensitivity and accuracy are relatively low. Data on the long term effect of training is missing in this dataset. Intuitively, job training would exert its influence on people's work in the long run. However, it can not be verified in this analysis. There is an imbalance in the number of participants who joined the training program.

Appendix I (Part1 R Code)

```
###### Clear environment and load libraries
rm(list = ls())
library(ggplot2)
library(rms)
library(MASS)
library(arm)
library(gganimate)
library(gifski)
library(av)
library(dplyr)
theme_set(theme_bw())

###### Load the data
lalondedata <-
  read.table(
    "lalondedata.txt",
    header = TRUE,
    sep = ",",
    colClasses = c(
      "factor",
      "factor",
      "numeric",
      "numeric",
      "factor",
      "factor",
      "factor",
      "factor",
      "numeric",
      "numeric",
      "numeric"
    )
  )

subrace <- lalondedata[c("black", "hispan")]
lalondedata$race <- apply(subrace, 1, function(x) {
  ifelse(x[1] == 1 & x[2] == 0, 1,
         ifelse(x[1] == 0 & x[2] == 1, 2, 0))
})
lalondedata$race <-
  factor(
    lalondedata$race,
    levels = c(0, 1, 2),
    labels = c("Otherwise", "Black", "Hispanic")
  )
lalondedata$diff <- lalondedata$re78 - lalondedata$re74
dim(lalondedata)
str(lalondedata)
summary(lalondedata)

###### EDA
# hist(lalondedata$re78 - lalondedata$re74, breaks = 20)
df1 <- data.frame(x=lalondedata$re78, time = 1)
```

```
df2 <- data.frame(x=lalondedata$diff, time = 2)
df <- cbind(df1, df2)

ggplot(df, aes(x = x)) +
  geom_histogram(
    aes(y = ..density..),
    color = "black",
    linetype = "dashed",
    fill = rainbow(25),
    binwidth = 2500
  ) +
  geom_density(alpha = .25, fill = "lightblue") +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Distribution of Real Annual Earnings in 1978",
       x = "Real Annual Earnings Difference") +
  theme_classic() + theme(legend.position = "none") +
  transition_states(time) +
  shadow_mark()

ggplot(lalondedata, aes(x = diff)) +
  geom_histogram(
    aes(y = ..density..),
    color = "black",
    linetype = "dashed",
    fill = rainbow(35),
    binwidth = 2500
  ) +
  geom_density(alpha = .25, fill = "lightblue") +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Distribution of Real Annual Earnings Difference between 1978 and 1974",
       x = "Real Annual Earnings Difference") +
  theme_classic() + theme(legend.position = "none")

# relationship b/w diff & each predictor
# age
ggplot(lalondedata, aes(x = age, y = diff)) +
  geom_point(alpha = .5, colour = "blue4") +
  geom_smooth(method = "lm", col = "red3") +
  theme_classic() +
  labs(title = "Difference in Earnings vs Age", x = "Age",
       y = "Difference in Earnings")

# educ
ggplot(lalondedata, aes(x = educ, y = diff)) +
  geom_point(alpha = .5, colour = "blue4") +
  geom_smooth(col = "red3") +
  theme_classic() +
  labs(title = "Difference in Earnings vs Education", x = "Education",
       y = "Difference in Earnings")

# treat
ggplot(lalondedata, aes(x = treat, y = diff, fill = treat)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Difference in Earnings vs Received Job Training", x = "Received Job Training",
       y = "Difference in Earnings") +
  theme_classic() + theme(legend.position = "none")

# race
ggplot(lalondedata, aes(x = race, y = diff, fill = race)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Difference in Earnings vs Races", x = "Races",
       y = "Difference in Earnings") +
  theme_classic() + theme(legend.position = "none")

# married
ggplot(lalondedata, aes(x = married, y = diff, fill = married)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Difference in Earnings vs Marital Status", x = "Marital Status",
       y = "Difference in Earnings") +
  theme_classic() + theme(legend.position = "none")

# nodegree
ggplot(lalondedata, aes(x = nodegree, y = diff, fill = nodegree)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Difference in Earnings vs No High School Degree", x = "No High School Degree",
       y = "Difference in Earnings") +
  theme_classic() + theme(legend.position = "none")

# interactions with age
ggplot(lalondedata, aes(x = age, y = diff)) +
  geom_point(alpha = .5, colour = "blue4") +
  geom_smooth(method = "lm", col = "red3") +
  theme_classic() +
  labs(title = "Difference in Earnings vs Age", x = "Age",
       y = "Difference in Earnings") +
  facet_wrap( ~ treat)

ggplot(lalondedata, aes(x = age, y = diff)) +
  geom_point(alpha = .5, colour = "blue4") +
  geom_smooth(method = "lm", col = "red3") +
  theme_classic() +
  labs(title = "Difference in Earnings vs Age", x = "Age",
       y = "Difference in Earnings") +
  facet_wrap( ~ race)

ggplot(lalondedata, aes(x = age, y = diff)) +
  geom_point(alpha = .5, colour = "blue4") +
  geom_smooth(method = "lm", col = "red3") +
  theme_classic() +
  labs(title = "Difference in Earnings vs Age", x = "Age",
       y = "Difference in Earnings") +
  facet_wrap( ~ married)

ggplot(lalondedata, aes(x = age, y = diff)) +
  geom_point(alpha = .5, colour = "blue4") +
  geom_smooth(method = "lm", col = "red3") +
  theme_classic() +
  labs(title = "Difference in Earnings vs Age", x = "Age",
       y = "Difference in Earnings") +
  facet_wrap( ~ nodegree)

# interactions with educ
ggplot(lalondedata, aes(x = educ, y = diff)) +
  geom_point(alpha = .5, colour = "blue4") +
  geom_smooth(method = "lm", col = "red3") +
  theme_classic() +
  labs(title = "Difference in Earnings vs Education", x = "Education",
       y = "Difference in Earnings") +
  facet_wrap( ~ treat)

ggplot(lalondedata, aes(x = educ, y = diff)) +
  geom_point(alpha = .5, colour = "blue4") +
  geom_smooth(method = "lm", col = "red3") +
  theme_classic() +
  labs(title = "Difference in Earnings vs Education", x = "Education",
       y = "Difference in Earnings") +
  facet_wrap( ~ race)

ggplot(lalondedata, aes(x = educ, y = diff)) +
  geom_point(alpha = .5, colour = "blue4") +
  geom_smooth(method = "lm", col = "red3") +
  theme_classic() +
  labs(title = "Difference in Earnings vs Education", x = "Education",
       y = "Difference in Earnings") +
  facet_wrap( ~ married)

ggplot(lalondedata, aes(x = educ, y = diff)) +
  geom_point(alpha = .5, colour = "blue4") +
  geom_smooth(method = "lm", col = "red3") +
  theme_classic() +
  labs(title = "Difference in Earnings vs Education", x = "Education",
       y = "Difference in Earnings") +
  facet_wrap( ~ nodegree)

# interactions with treat
ggplot(lalondedata, aes(x = treat, y = diff, fill = treat)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Difference in Earnings vs Received Job Training", x = "Received Job Training",
       y = "Difference in Earnings") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap( ~ race)

ggplot(lalondedata, aes(x = treat, y = diff, fill = treat)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Difference in Earnings vs Received Job Training", x = "Received Job Training",
       y = "Difference in Earnings") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap( ~ married)

ggplot(lalondedata, aes(x = treat, y = diff, fill = treat)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Difference in Earnings vs Received Job Training", x = "Received Job Training",
       y = "Difference in Earnings") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap( ~ nodegree)

# interactions with race
ggplot(lalondedata, aes(x = race, y = diff, fill = race)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Difference in Earnings vs Races", x = "Races",
       y = "Difference in Earnings") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap( ~ married)

ggplot(lalondedata, aes(x = race, y = diff, fill = race)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Difference in Earnings vs Races", x = "Races",
       y = "Difference in Earnings") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap( ~ nodegree)

# interactions with married
ggplot(lalondedata, aes(x = married, y = diff, fill = married)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Difference in Earnings vs Marital Status", x = "Marital Status",
       y = "Difference in Earnings") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap( ~ nodegree)

###### Modeling Fitting
# mean center the numerical predictors
lalondedata$agec <- c(scale(lalondedata$age, scale = F))
lalondedata$educc <- c(scale(lalondedata$educ, scale = F))
lalondedata$educc2 <- lalondedata$educc ^ 2

# Null Model
Model_Null <- lm(diff ~ treat * race, data = lalondedata)
summary(Model_Null)

# Full Model
Model_Full <-
  lm(diff ~ (agec + educc + treat + race + married + nodegree) ^ 2 + educc2,
     data = lalondedata)
summary(Model_Full)

# Stepwise
```

```
Model_stepwise_aic <-
  step(Model_Null,
       scope = Model_Full,
       direction = "both",
       trace = 0)
summary(Model_stepwise_aic)

Model_forward_aic <-
  step(Model_Null,
       scope = Model_Full,
       direction = "forward",
       trace = 0)
summary(Model_forward_aic)

Model_backward_aic <-
  step(Model_Null,
       scope = Model_Full,
       direction = "backward",
       trace = 0)
summary(Model_backward_aic)


## Final model is Model2
Model2 <-
  lm(
    diff ~ treat + black + hispan + agec + married + treat:agec + educc + educc2
    + agec:married,
    data = lalondedata
  )

summary(Model2)

confint(Model2, level = 0.95)

##### Model Assesment
vif(Model2)

# Assumptions
plot(Model2, which = 1:5, col = c("blue4"))

ggplot(lalondedata, aes(x = agec, y = Model2$residuals)) +
  geom_point(alpha = .7) + geom_hline(yintercept = 0, col = "red3") + theme_classic() +
  labs(title = "Residuals vs Age (Centered)", x = "Age (Centered)", y =
         "Residuals")
ggplot(lalondedata, aes(x = educc, y = Model2$residuals)) +
  geom_point(alpha = .7) + geom_hline(yintercept = 0, col = "red3") + theme_classic() +
  labs(title = "Residuals vs Education (Centered)", x = "Education (Centered)", y =
         "Residuals")
```

Appendix II (Part2 R Code)

```
###### Clear environment and load libraries
rm(list = ls())
library(ggplot2)
library(rms)
library(MASS)
library(arm)
library(pROC)
library(e1071)
library(caret)
library(dplyr)
library(tidyr)
require(gridExtra)

###### Load the data
lalondedata <-
  read.table(
    "lalondedata.txt",
    header = TRUE,
    sep = ",",
    colClasses = c(
      "factor",
      "factor",
      "numeric",
      "numeric",
      "factor",
      "factor",
      "factor",
      "factor",
      "numeric",
      "numeric",
      "numeric"
    )
  )

lalondedata$earn <- ifelse(lalondedata$re78 > 0, 1, 0)

lalondedata$earnf <-
  factor(
    ifelse(lalondedata$re78 > 0, 1, 0),
    levels = c(0, 1),
    labels = c("Zero", "Positive")
  )

###### Exploratory data analysis
# earn vs age
ggplot(lalondedata, aes(x = earnf, y = age, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Age",
       x = "Had salaries or no?", y = "Age") +
  theme_classic() + theme(legend.position = "none")
```

```
# earn vs age by treat
ggplot(lalondedata, aes(x = earnf, y = age, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Age by Treat",
       x = "Had salaries or no?", y = "Age") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap(~ treat)

# earn vs age by black
ggplot(lalondedata, aes(x = earnf, y = age, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Age by Black Race",
       x = "Had salaries or no?", y = "Age") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap(~ black)

# earn vs age by hispan
ggplot(lalondedata, aes(x = earnf, y = age, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Age by Hispanic Ethinicity",
       x = "Had salaries or no?", y = "Age") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap(~ hispan)

# earn vs age by married
ggplot(lalondedata, aes(x = earnf, y = age, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Age by Marital Status",
       x = "Had salaries or no?", y = "Age") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap(~ married)

# earn vs age by nodegree
ggplot(lalondedata, aes(x = earnf, y = age, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Age by High school degree",
       x = "Had salaries or no?", y = "Age") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap(~ nodegree)

# earn vs educ
ggplot(lalondedata, aes(x = earnf, y = educ, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Education",
       x = "Had salaries or no?", y = "Education") +
  theme_classic() + theme(legend.position = "none")

# earn vs educ by treat
```

```
ggplot(lalondedata, aes(x = earnf, y = educ, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Education by Treat",
       x = "Had salaries or no?", y = "Education") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap(~ treat)

# earn vs educ by black
ggplot(lalondedata, aes(x = earnf, y = educ, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Education by Black Race",
       x = "Had salaries or no?", y = "Education") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap(~ black)

# earn vs educ by hispan
ggplot(lalondedata, aes(x = earnf, y = educ, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Education by Hispanic Ethinicity",
       x = "Had salaries or no?", y = "Education") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap(~ hispan)

# earn vs educ by married
ggplot(lalondedata, aes(x = earnf, y = educ, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Education by Marital Status",
       x = "Had salaries or no?", y = "Education") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap(~ married)

# earn vs educ by nodegree
ggplot(lalondedata, aes(x = earnf, y = educ, fill = earnf)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette = "Reds") +
  labs(title = "Had salaries or not vs Education by High school degree",
       x = "Had salaries or no?", y = "Education") +
  theme_classic() + theme(legend.position = "none") +
  facet_wrap(~ nodegree)

# earn vs treat
t1 <-
  round(apply(table(lalondedata[, c("earnf", "treat")]) /
                sum(table(lalondedata[, c("earnf", "treat")])),
              2, function(x)
                x / sum(x)), 4)
t2 <-
  round(apply(table(lalondedata[, c("earnf", "black")]) /
                sum(table(lalondedata[, c("earnf", "black")])),
              2, function(x)
              x / sum(x)), 4)
t3 <-
  round(apply(table(lalondedata[, c("earnf", "hispan")]) /
              sum(table(lalondedata[, c("earnf", "hispan")])),
            2, function(x)
              x / sum(x)), 4)
t4 <-
  round(apply(table(lalondedata[, c("earnf", "married")]) /
              sum(table(lalondedata[, c("earnf", "married")])),
            2, function(x)
              x / sum(x)), 4)
t5 <-
  round(apply(table(lalondedata[, c("earnf", "nodegree")]) /
              sum(table(lalondedata[, c("earnf", "nodegree")])),
            2, function(x)
              x / sum(x)), 4)

chitest1 <- chisq.test(table(lalondedata[, c("earnf", "treat")]))
chitest2 <- chisq.test(table(lalondedata[, c("earnf", "black")]))
chitest3 <- chisq.test(table(lalondedata[, c("earnf", "hispan")]))
chitest4 <- chisq.test(table(lalondedata[, c("earnf", "married")]))
chitest5 <- chisq.test(table(lalondedata[, c("earnf", "nodegree")]))

black0 <- lalondedata %>%
  filter(black == 0)
black1 <- lalondedata %>%
  filter(black == 1)
apply(table(black0[, c("earnf", "treat")]) / sum(table(black0[, c("earnf", "treat")])),
      2, function(x)
        x / sum(x))
apply(table(black1[, c("earnf", "treat")]) / sum(table(black1[, c("earnf", "treat")])),
      2, function(x)
        x / sum(x))
# + black:treat

hispan0 <- lalondedata %>%
  filter(hispan == 0)
hispan1 <- lalondedata %>%
  filter(hispan == 1)
apply(table(hispan0[, c("earnf", "treat")]) / sum(table(hispan0[, c("earnf", "treat")])),
      2, function(x)
        x / sum(x))
apply(table(hispan1[, c("earnf", "treat")]) / sum(table(hispan1[, c("earnf", "treat")])),
      2, function(x)
        x / sum(x))
# + treat:hispan

married0 <- lalondedata %>%
  filter(married == 0)
married1 <- lalondedata %>%
  filter(married == 1)
apply(table(married0[, c("earnf", "treat")]) / sum(table(married0[, c("earnf", "treat")])),
      2, function(x)
        x / sum(x))
```

```
apply(table(married1[, c("earnf", "treat")]) / sum(table(married1[, c("earnf", "treat")])),
      2, function(x)
        x / sum(x))
# ~ treat:married

nodegree0 <- lalondedata %>%
  filter(nodegree == 0)
nodegree1 <- lalondedata %>%
  filter(nodegree == 1)
apply(table(nodegree0[, c("earnf", "treat")]) / sum(table(nodegree0[, c("earnf", "treat")])),
      2, function(x)
        x / sum(x))
apply(table(nodegree1[, c("earnf", "treat")]) / sum(table(nodegree1[, c("earnf", "treat")])),
      2, function(x)
        x / sum(x))
# - treat:nodegree

black0 <- lalondedata %>%
  filter(black == 0)
black1 <- lalondedata %>%
  filter(black == 1)
apply(table(black0[, c("earnf", "married")]) / sum(table(black0[, c("earnf", "married")])),
      2, function(x)
        x / sum(x))
apply(table(black1[, c("earnf", "married")]) / sum(table(black1[, c("earnf", "married")])),
      2, function(x)
        x / sum(x))
# + married:black

black0 <- lalondedata %>%
  filter(black == 0)
black1 <- lalondedata %>%
  filter(black == 1)
apply(table(black0[, c("earnf", "nodegree")]) / sum(table(black0[, c("earnf", "nodegree")])),
      2, function(x)
        x / sum(x))
apply(table(black1[, c("earnf", "nodegree")]) / sum(table(black1[, c("earnf", "nodegree")])),
      2, function(x)
        x / sum(x))
# + nodegree:black

hispan0 <- lalondedata %>%
  filter(hispan == 0)
hispan1 <- lalondedata %>%
  filter(hispan == 1)
apply(table(hispan0[, c("earnf", "nodegree")]) / sum(table(hispan0[, c("earnf", "nodegree")])),
      2, function(x)
        x / sum(x))
apply(table(hispan1[, c("earnf", "nodegree")]) / sum(table(hispan1[, c("earnf", "nodegree")])),
      2, function(x)
        x / sum(x))
# + nodegree:hispan

hispan0 <- lalondedata %>%
  filter(hispan == 0)
hispan1 <- lalondedata %>%
  filter(hispan == 1)
apply(table(hispan0[, c("earnf", "married")]) / sum(table(hispan0[, c("earnf", "married")])),
      2, function(x)
        x / sum(x))
apply(table(hispan1[, c("earnf", "married")]) / sum(table(hispan1[, c("earnf", "married")])),
      2, function(x)
        x / sum(x))
# + hispan:married

nodegree0 <- lalondedata %>%
  filter(nodegree == 0)
nodegree1 <- lalondedata %>%
  filter(nodegree == 1)
apply(table(nodegree0[, c("earnf", "married")]) / sum(table(nodegree0[, c("earnf", "married")])),
      2, function(x)
        x / sum(x))
apply(table(nodegree1[, c("earnf", "married")]) / sum(table(nodegree1[, c("earnf", "married")])),
      2, function(x)
        x / sum(x))
# + nodegree:married

#binned plots
par(mfcol = c(1, 1))
binnedplot(
  x = lalondedata$age,
  y = lalondedata$earn,
  xlab = "Age",
  ylim = c(0, 1),
  col.pts = "navy",
  ylab = "Had salaries or not ",
  main = "Binned Plot for Had salaries or not  w.r.t \nAge",
  col.int = "white"
)

binnedplot(
  x = lalondedata$educ,
  y = lalondedata$earn,
  xlab = "Education",
  ylim = c(0, 1),
  col.pts = "navy",
  ylab = "Had salaries or not ",
  main = "Binned Plot for Had salaries or not  w.r.t \nEducation",
  col.int = "white"
)

###### Model fitting
lalondedata$agec <- lalondedata$age - mean(lalondedata$age)
lalondedata$agec2 <- lalondedata$agec ^ 2
lalondedata$educc <- lalondedata$educ - mean(lalondedata$educ)

ModelNull <-
  glm(earn ~ treat + black + agec + agec2,
      data = lalondedata,
      family = binomial)
summary(ModelNull)

ModelFull <-
  glm(
    earn ~ (agec + educc + treat + black + hispan + married + nodegree) ^ 2 + agec2,
    data = lalondedata,
    family = binomial
  )
summary(ModelFull)

Model_stepwise_aic <- step(ModelNull,
                           scope = ModelFull,
                           direction = "both",
                           trace = 0)
summary(Model_stepwise_aic)

Model_forward_aic <- step(ModelNull,
                          scope = ModelFull,
                          direction = "forward",
                          trace = 0)
summary(Model_forward_aic)

Model_backward_aic <- step(ModelNull,
                           scope = ModelFull,
                           direction = "backward",
                           trace = 0)
summary(Model_backward_aic)

Model1 <-
  glm(earn ~ treat + black + agec + agec2 + agec:treat,
      data = lalondedata,
      family = binomial)
summary(Model1)
anova(ModelNull, Model1, test = "Chisq")

Model2 <-
  glm(
    earn ~ treat + black + agec + agec2 + agec:treat + educ + educ:black,
    data = lalondedata,
    family = binomial
  )
summary(Model2)
anova(Model1, Model2, test = "Chisq")

Model3 <-
  glm(
    earn ~ treat + black + agec + agec2 + agec:treat + hispan + hispan:educ,
    data = lalondedata,
    family = binomial
  )
summary(Model3)
anova(Model1, Model3, test = "Chisq")
```

```
Model4 <-
  glm(
    earn ~ treat + black + agec + agec2 + agec:treat +  married + married:educc,
    data = lalondedata,
    family = binomial
  )
summary(Model4)
anova(Model1, Model4, test = "Chisq")

Model5 <-
  glm(
    earn ~ treat + black + agec + agec2 + agec:treat +  nodegree + nodegree:agec,
    data = lalondedata,
    family = binomial
  )
summary(Model5)
anova(Model1, Model5, test = "Chisq")

Model6 <-
  glm(
    earn ~ treat + black + agec + agec2 + agec:treat +  nodegree + nodegree:agec + nodegree:educ,
    data = lalondedata,
    family = binomial
  )
summary(Model6)
anova(Model1, Model6, test = "Chisq")

Model7 <-
  glm(
    earn ~ earn ~ treat + black + agec + agec2 + agec:treat +  nodegree + nodegree:agec +
      treat:black + treat:hispan + treat:married + treat:nodegree +
      black:hispan + black:married + black:nodegree +
      hispan:married + hispan:nodegree +
      nodegree:married + hispan + married,
    data = lalondedata,
    family = binomial
  )
summary(Model7)
anova(Model6, Model7, test = "Chisq")

Model8 <- step(Model5,
               scope = Model7,
               direction = "both",
               trace = 0)
summary(Model8)

Model9 <- step(Model5,
               scope = Model7,
               direction = "forward",
               trace = 0)
summary(Model9)

Model10 <- step(Model5,
                scope = Model7,
                direction = "backward",
                trace = 0)
summary(Model10)

Model11 <-
  glm(
    earn ~ treat + black + agec + agec2 + agec:treat +  nodegree + nodegree:agec +
      hispan + married + hispan:married,
    data = lalondedata,
    family = binomial
  )
summary(Model11)
anova(Model5, Model11, test = "Chisq")

Model12 <-
  glm(
    earn ~ treat + black + agec + agec2 + agec:treat +
      hispan + married + hispan:married,
    data = lalondedata,
    family = binomial
  )
summary(Model12)
anova(Model12, Model11, test = "Chisq")

FinalModel <- Model11
summary(FinalModel)

Model13 <-  glm(
  earn ~ treat + race + agec + agec2 + agec:treat++married + race:married,
  data = lalondedata,
  family = binomial
)
summary(Model13)

Model14 <- glm(
  earn ~ treat + black + agec + agec2 + hispan + married + agec:treat +
    nodegree + nodegree:agec + hispan:married + black:treat,
  data = lalondedata,
  family = binomial
)
summary(Model14)
anova(Model11, Model14, test = "Chisq")
# - black:treat

Model15 <- glm(
  earn ~ treat + black + agec + agec2 + hispan + married + agec:treat +
    nodegree + nodegree:agec + hispan:married + treat:hispan,
  data = lalondedata,
  family = binomial
)
summary(Model15)
anova(Model11, Model15, test = "Chisq")
# + treat:hispan
```

```
Model16 <- glm(
  earn ~ treat + black + agec + agec2 + hispan + married + agec:treat +
    nodegree + nodegree:agec + hispan:married  + treat:hispan + treat:married,
  data = lalondedata,
  family = binomial
)
summary(Model16)
anova(Model15, Model16, test = "Chisq")

Model17 <- glm(
  earn ~ treat + black + agec + agec2 + hispan + married + agec:treat +
    nodegree + nodegree:agec + hispan:married  + treat:hispan + married:black,
  data = lalondedata,
  family = binomial
)
summary(Model17)
anova(Model15, Model17, test = "Chisq")

Model18 <- glm(
  earn ~ treat + black + agec + agec2 + hispan + married + agec:treat +
    nodegree + nodegree:agec + hispan:married  + treat:hispan + nodegree:black,
  data = lalondedata,
  family = binomial
)
summary(Model18)
anova(Model15, Model18, test = "Chisq")

Model19 <- glm(
  earn ~ treat + black + agec + agec2 + hispan + married + agec:treat +
    nodegree + nodegree:agec + hispan:married  + treat:hispan + nodegree:hispan,
  data = lalondedata,
  family = binomial
)
summary(Model19)
anova(Model15, Model19, test = "Chisq")

Model20 <- glm(
  earn ~ treat + black + agec + agec2 + hispan + married + agec:treat +
    nodegree + nodegree:agec + hispan:married  + treat:hispan + nodegree:married,
  data = lalondedata,
  family = binomial
)
summary(Model20)
anova(Model15, Model20, test = "Chisq")

anova(Model11, Model15, test = "Chisq")

Model21 <- step(Model11,
                scope = Model15,
                direction = "both",
                trace = 0)
summary(Model21)
# Model21 = Model11
```

```
FinalModel <- Model11
summary(FinalModel)


###### Model fitting
rawresid <- residuals(FinalModel, "resp")

#binned residual plots
par(mfrow = c(1, 1))
binnedplot(
  x = fitted(FinalModel),
  y = rawresid,
  xlab = "Pred. probabilities",
  col.int = "red4",
  ylab = "Avg. residuals",
  main = "Binned residual plot",
  col.pts = "navy"
)

binnedplot(
  x = lalondedata$agec,
  y = rawresid,
  xlab = "Age (Centered)",
  col.int = "red4",
  ylab = "Avg. residuals",
  main = "Binned residual plot",
  col.pts = "navy"
)

binnedplot(
  x = lalondedata$educc,
  y = rawresid,
  xlab = "Education (Centered)",
  col.int = "red4",
  ylab = "Avg. residuals",
  main = "Binned residual plot",
  col.pts = "navy"
)


######## Model Validation
Conf_mat <-
  confusionMatrix(as.factor(ifelse(
    fitted(FinalModel) >= mean(lalondedata$earn), "1", "0"
  )),
  as.factor(lalondedata$earn), positive = "1")
Conf_mat$table
Conf_mat$overall["Accuracy"]
Conf_mat$byClass[c("Sensitivity", "Specificity")]

roc(
  lalondedata$earn,
  fitted(FinalModel),
  plot = T,
  print.thres = "best",
  legacy.axes = T,
  print.auc = T,
  col = "red3",
  quiet = TRUE
)

###### Confidence Interval
confint(FinalModel)
```