Team Project 1

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

Summary

Through this report we try to analyze the impact of a training program by gauging its effect on the real annual earnings and through its likeliness to help the participants earn a non-zero wage. To do so, we use linear and logitic regression models respectively. We use a dataset from the National Supported Work (NSW) Demonstration (1975-1978) for the same and build models using the AIC/BIC criterion based on the statistical significance and reasonableness. We then find out that ..... have significant association with annual incomes suggesting that taking the training might be an important factor associated with increase in wages. Also, We find that non-zero wages are associated with ..... suggesting that training is not a very significant factor associated with non-zero wages. Thus, it can be understood that training program as a factor does not seem to have a very strong association with improvement in wages or by providing a source of earning.

For detailed information on this research, please check the following papers.

- Paper 1
- Paper 2

Introduction

In the 1970s, researchers in the United States ran several randomized experiments to evaluate public policy programs. One of the most famous experiments is the National Supported Work (NSW) Demonstration, in which researchers wanted to assess whether or not job training for disadvantaged workers had an effect on their wages. Based on a subset of the investigation, in order to undersatnd 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 look at quantify the effect of the treatment, that is, 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 if there are other interesting associations with wages.

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 the 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, tries building a model by exploring models built with the help of AIC and BIC criteria using

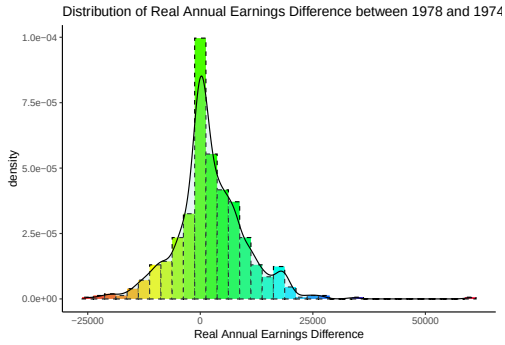
forward and stepwise model building and chooses the most suitable model on the basis of accuracy and plausibility to answer the above questions.

PART I

DATA & EDA

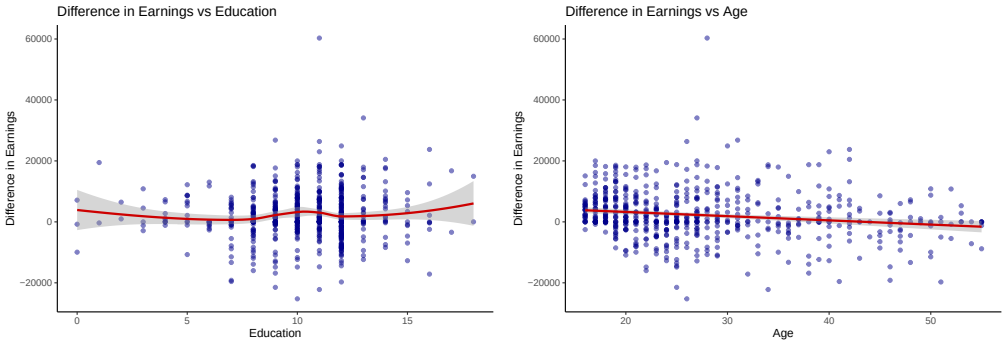
Response Variable

Since the goal is to determine the effects of job training on salary increment, the response variable is the difference in salary between 1974 and 1978. Its distribution is quite normal; hence there is no transformation for the response variable.



Predict Vraibles

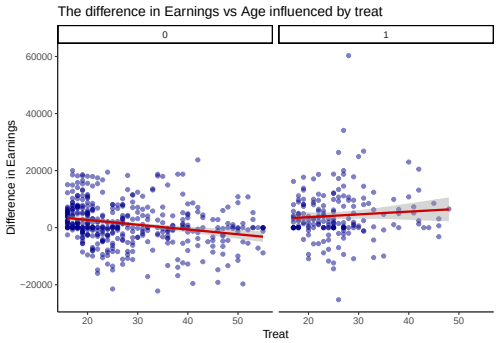
Unlike the relationship between age and 'treat' (Right), the relationship between education(left) and 'treat' is not linear. It indicates that there is some non-linear transformation should be performed on education. After evaluation, we decide to use the square to transform the variable, and the final model's p-value confirmed the is.



Multicollinearity

Intuitively, education duration has a strong correlation with a high school degree. In this dataset, the correlation of the two variables is -0.7 which suggests that we could not include both of them in our model. After evaluation, education duration was preserved since it can provide more information than the high school degree variable.

Interactions



- 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

1. We find some signs of interaction between "treat" and "age" in EDA. The business was confirmed in our model. Its p-value is significantly small than 0.05.
2. Interaction between "married" and "age" was found in this step. It's p-Value is slightly above 0.05, but it small than 0.1. The ANOVA test also indicates that it would improve our model significantly. Hence it was preserved in our final model.
3. The education was not linear in the plot of EDA, and this finding was also confirmed. Its square transformation was significant (p-value: 0.03).

Final model

$$\begin{aligned} diff_i = &\beta_0 + \beta_1 * black_i + \beta_2 * hispan_i + \beta_3 * agec_i + \\ &\beta_4 * married_i + \beta_5 * treat_i : agec_i + \beta_6 * educc_i + \\ &\beta_7 * educc_i^2 + \beta_8 agec : married_i + \varepsilon_i \end{aligned}$$

- agec: Centred age
- educc: Centred educ

Model Summary & CI

	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

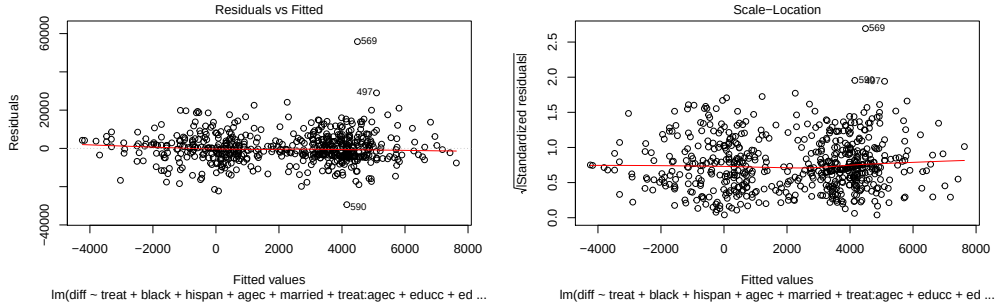
Table 2: Evaluationl

	2.5 %	97.5 %
(Intercept)	455.53	3031.80
treat1	1516.63	4991.57
black1	-2374.90	977.70
hispan1	-1433.19	2679.78
agec	-322.99	-117.20
married1	-3307.36	-451.84
educc	-107.05	409.54
educc2	3.16	107.81
treat1:agec	124.20	476.44
agec:married1	-1.69	277.68

Table 3: Confidence Interval

Model Verification

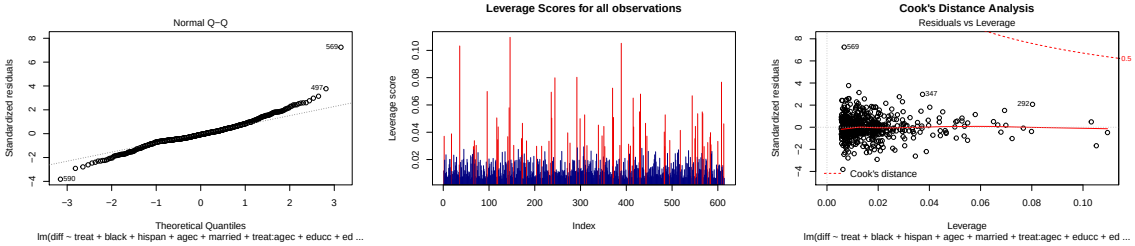
Residuals



- The residuals are scattered randomly; there is no apparent trend in the plots.
- The error is no correlation of error terms in the plot.
- The variance of the error is constant, there is no apparent change along the x-axis.

Summary: According to residual analysis, there is no obvious evidence indicate the assumptions of linear regression were broken.

Outliers and High Leverage



- There are a few outliers under this model.
- There are some high leverage points.
- According to cook's distance, there is no high influence points (> 0.5).

Summary: There are some outliers and high leverage points; however, there are not high influence data. Hence, these data points can be preserved in the model without worry.

Collinearity

	names	x
1	treat1	1.69
2	black1	1.79
3	hispan1	1.17
4	agec	2.75
5	married1	1.32
6	educc	1.23
7	educc2	1.26
8	treat1:agec	1.31
9	agec:married1	2.25

Table 4: VIF

- According to VIF table, there is obvious colineary problem in this model.

Conclusion

1. Tread has positive effects on workers' annual salary because its p-value is significant. Controlling other factors, taking job training would increase $3254 on annual salary on average. It's 95% CI is (1516, 4991)
2. The effect varies by age. 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.
3. Other interesting associations with wages:

- Marriage would significantly bring down workers' annual salary by $1879 (95% CI is 452, 3307)
- Education duration would increase workers' salaries. For 1 unit increase for its square, the annual salary would increase by $55 (95% CI: 3, 108)
- Age and married have weak interaction. Controlled other factors, for the married workers, while the old ones would receive more salary. One year increase on age would raise the workers' salary by $137 per year (95% CI: -2, 278)

Deficiency

1. The final model's R-squared is only 0.088, which is relatively low.
2. Some outliers deserve further investigation.

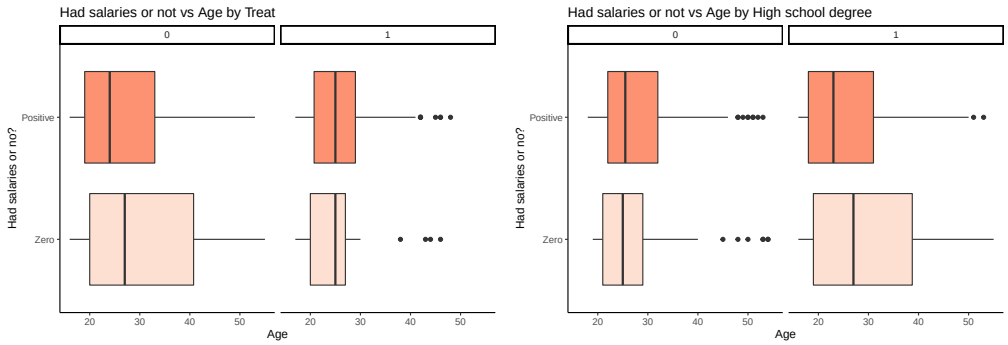
Part II

DATA & EDA

Collinearity

- As the same reason mentioned in Part I, we exclude "educ" from our model.

Interactions



- As reported by the first plots (left), the role of age in the employment rate is different according to whether the works take job training or not.
- The second plot points out that the role of age in employment rate is also influenced by whether the works have a high school degree or not.

	0	1
Zero	0.26	0.21
Positive	0.74	0.79

Table 5: Employed rate with marital status - Non Hispanic

	0	1
Zero	0.08	0.28
Positive	0.92	0.72

Table 6: Employed rate with marital status - Hispanic

- The two tables show some differences in the relationship between marital status and employment rate according to race. For the Hispanic works, the single workers' unemployment rate is obviously 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.

Model

Model selection

1. On the basis of the above, the logistic regression model is developed after considering the stepwise and forward built models using the AIC criteria and then selected. The Null Model used is :

$$logit(\pi_i) = \beta_0 + \beta_1 * treat_i + \beta_2 * black_i + \beta_3 * agec_i + \beta_4 * agec_i^2 + \varepsilon_i$$

and the Full Model used is:

$$logit(\pi_i) = \beta_0 + (\beta_1 agec_i + \beta_2 educc_i + \beta_3 treat_i + \beta_4 black_i + \beta_5 hispan_i + \beta_6 married_i + \beta_7 nodegree_i)^2 + \beta_8 * age_i c^2 + \varepsilon_i$$

2. By AIC forward-searching, we find a square transform of "age" is significant in the model. After verification with ANOVA test, we decide to keep this transformation in the final model
3. The interactions between "age" and "married" and "degree" were confirmed by the p-value. Even though the p-value is "married" and "treat" not strong significant, the ANOVA test tips it means full.
4. The interaction between "hispan1" and "married" was confirmed, and its p-value is significant.
5. Tread was not significant in our final model. However, it is a key predictor variable for answering the question. Hence, we preserve it in our final model.

Final model

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

- agec: Centred age

Model Summary & CI

	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 7: Coefficients

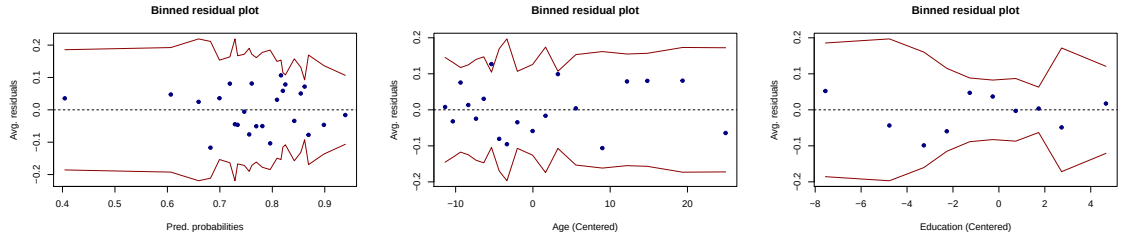
	2.5 %	97.5 %
(Intercept)	0.94	1.98
treat1	-0.32	0.80
black1	-1.06	-0.02
agec	-0.03	0.06
agec2	-0.00	0.00
nodegree1	-0.44	0.40
hispan1	0.08	2.68
married1	-0.15	0.88
agec:nodegree1	-0.09	-0.00
hispan1:married1	-3.47	-0.37
treat1:agec	-0.01	0.10

Table 8: Confidence Interval

	NULL_deviance	Residual_deviance
1	666.50	627.54

Table 9: Deviance

Model Verification



Residuals

- According to the binnedplots, about 95% points reside inside the red bend. It is a strong justification for the model's efficiency.

Deviance

- Null model deviance : 666.5
- Final model deviance : 627.54
- The decrease of deviance indicates that the model is valid.

Outliers and High Leverage

	names	x
1	treat1	1.82
2	black1	1.79
3	agec	6.83
4	agec2	3.06
5	nodegree1	1.09
6	hispan1	3.35
7	married1	1.69
8	agec:nodegree1	3.54
9	hispan1:married1	3.44
10	treat1:agec	1.38

Table 10: VIF

Collinearity

- According to the table, all item's VIF value are below 10. Hence, there is no remarkable collinearity in this model.

Summary Various indicators point out that the final model is valid and can be used to answer the questions about the dataset.

Model Assessment

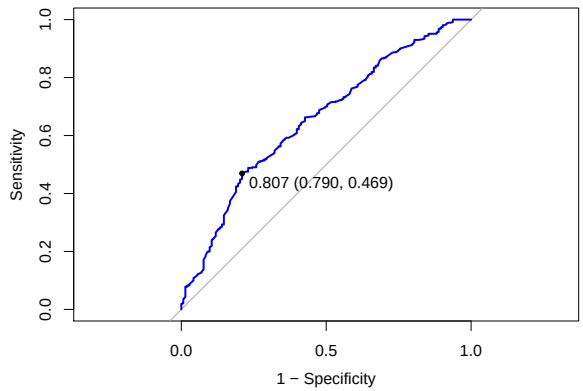
- Sensitivity : 0.59
- Specificity : 0.64

	0	1
0	91	192
1	52	279

Table 11: Confusion Matrix

- Accuracy : 0.6026059

ROC



- AUC: 0.65
- Cut-off: 0.81

Conclusion

- Since the treat's p-value is not significant, receiving job training would not directly influence the odds of these worker's positive wages. However, the interaction item "treat1:agec" is relatively significant.
- Some other factors would influence a worker's odds of getting positive wages.

1. Race of black. For a black worker, his/her odds of getting a job are 41% lower than a non-black worker under the same other conditions.
2. Race of Hispanic and marital status. Controlling other factors, a Hispanic worker's odd to be employed is 231% higher than others. However, the effect would bring down by being married. A married Hispanic worker's odds of getting a job are 45% lower than others under the same conditions.
3. Age. Age is a complex factor in our model. Generally speaking, aging would increase these workers' odds of being jobless. Taking the job training may slow down the trends.

Deficiency

- The effect of age is hard to interpret because it involves square transformation and interaction with treat. Moreover, their significants are not strong enough. This would cause the model to become less convincing.
- 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.

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 + educ + 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)
```