# Job Training Effects Analysis

Dapo Adegbile (Presenter) Christopher Oblak (Coordinator, Checker )
Sutianyi Wen (Programmer) Jiaman Betty Wu (Writer)

September 27, 2020

# PART I

***Summary***

In this analysis, we build a linear regression model to analyze how receiving job training affected the volume of money individuals earned. There was substantial evidence to suggest that individuals who received training had a positive impact of roughly $2415.00 over their control group counterpart while everything else held the same. Negatively impacting the likely hood of higher wages was independent variables of marriage and age. Where a marriage decreased the revenue amount by ~$1757, and for every year someone lived, their overall revenue would drop by ~$136 while everything else held the same.

***Introduction***

In the 1970s, researchers wanted to evaluate whether or not job training for disadvantaged workers had an effect on their wages. They collected data from workers who were randomly assigned either to receive training or not in the National Supported Work (NSW) Demonstration. For our analysis purpose, we use a subset of the original data. Throughout the analysis, we are motivated to find the effect that job training had on the overall revenue individuals were able to attain. Moreover, we would like to quantify the effect of job training on workers and find the range of the effects. Finally, we want to know if training effects differ by demographic groups and if there are other interesting relationships within data.

***Data***

**Exploratory Data Anlysis & Data Summary** The dataset used in the analysis contains 614 observations on 10 variables. Based on the research question of interest, we generate the response variable `wage_inc` by taking the difference between wage in 1978 and 1974 for each observation. This variable helps us to capture the wage increases between people who received training and those who did not. `wage_inc` ranges from -25257 to 60308 with the mean at 2235. Based on the histogram, the response variable roughly follows the normal distribution with. However, it is important to note that there are potential outliers. The 75th percentile is only 6460, however the maximum goes up to 60308. Therefore, this is important to consider potential outliers and influential points in model fitting later.

From the 10 variables, we collapse variables that relate to racial information into one variable `race`. The first interaction was Education as a factor vs Wages. When plotting these two variables we see that we have a number of perceived outliers on the boxplots. Additionally we could see that as the education of workers increase, the wages also increase, which reaffirms the thought that with more education comes higher wages. `race` takes the value "0" for non-Hispanic and non-Black observations, "1" for Hispanic, and "2" for Black. We constructed the `race` variable because it can help us to explore the how demographics might interact with other variables to affect wage. The majority (299 out of 614) of the sample are not Hispanic nor Black, followed by Black (243), and Hispanic (72). In the sample, Blacks have the largest wage increase (3177.6), followed by Hispanics (2675.1), and the remaining people (1363.6).

| | treat | age | married | wage_inc | race | educ_fac |
|---|---|---|---|---|---|---|
| | 0:429 | Min. :16.00 | 0:359 | Min. :-25257 | Neither:299 | Middle School:134 |
| | 1:185 | 1st Qu.:20.00 | 1:255 | 1st Qu.: -1240 | Hispan : 72 | High School :410 |
| | NA | Median :25.00 | NA | Median : 1262 | Black :243 | College : 70 |
| | NA | Mean :27.36 | NA | Mean : 2235 | NA | NA |
| | NA | 3rd Qu.:32.00 | NA | 3rd Qu.: 6460 | NA | NA |
| | NA | Max. :55.00 | NA | Max. : 60308 | NA | NA |

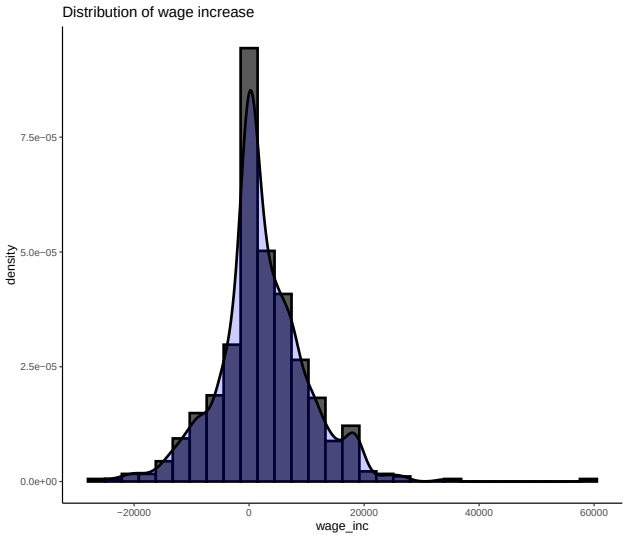
We first inspect the relationship between `wage_inc` and `educ` using a scatter plot, however, we do not find any discernible linear trend between the two variable. After researching more on the relationship between income and education, we learned about the *Sheepskin Effect*[1] which states that people possessing an academic degree have a greater income than people who have an equivalent amount of studying without possessing an academic degree. Inspired by the *Sheepskin Effect*, we construct a categorical variable `educ_fac` based on `educ`, a numeric variable that indicates the years of education. `educ_fac` indicates whether a person is have an education level that is middle school or lower (`educ` less than 9), high school (`educ` between 9 and 12), and college and above (`educ` more that 13). Most people (410) has high school education, 134 people have an education level that is middle school or lower, and 70 people have college education or higher. As shown in table, the wage increases more as people get higher education levels.

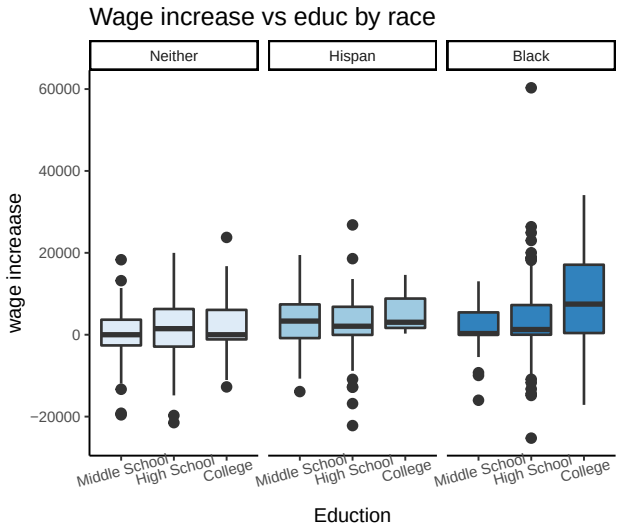
| educ_fac | wage_inc |
|---|---|
| Middle School | 1156.133 |
| High School | 2346.428 |
| College | 3650.138 |

Other variables include `treat`, `married`, and `age`. `treat` is indicates whether someone is in the treatment group. 429 observations are in the control group, the remaining 185 are in the treatment group. The average wage increase for the treatment group is 4253.6 and the average wage increase for the control group is 1364.9. `married` indicates whether an observation is married. Out of the 255 people who are married, the average wage increase is 403.6. The average wage increase for the remaining 429 unmarried people is 3536.3. In the sample, the age ranges from 16 to 55 years old, with an average at 27 years old.

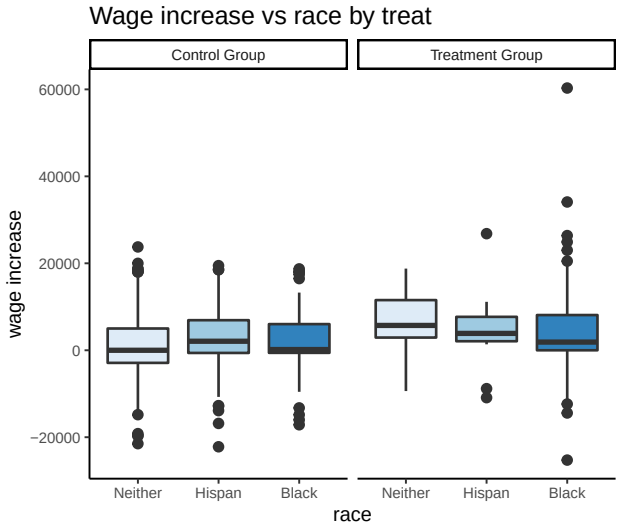
The plot "Wage increase vs educ by race" below shows the relationship between wage and education by race. We notice that for neither Black nor Hispanic, and Hispanic people there is not a consistent increase in the wage as a function of education. But with Black people, we see a consistent increase in wage as education increases. The could suggest that the effects of education on wage are different for Blacks compare to other groups. The plot "Wage vs age by treat" shows that for the control group, there is a weak negative linear trend between wage increase and age; however, the relationship is reversed for the treatment group. Furthermore, we also explore the relationship between treatment and wage increase by race. The plot "Wage increase vs race by treat" demonstrates that for the control group, the wage increase among all three races are similar with Hispanic slightly higher. However, in the treatment group, the wage increase for someone who is neither Hispanic nor Black is the greatest, followed by Hispanic, and Black. These differential effects displayed in the plots suggest that it is resonable to explore the relationship more by incorporating interaction terms for modeling.

[1] https://en.wikipedia.org/wiki/Sheepskin_effect








### *Model*

**Model Building** Our full model is given as below:

$$y_i = \boldsymbol{x_i}\boldsymbol{\beta} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

where $y_i$ is `wage_inc`, the independent variables $\boldsymbol{x_i}$ includes includes main effect variables: `treat`, `agec`, `agec_sq`, `educ_fac`, `race`, `married` where`agec` and `agec_sq` are mean-centered age and squared mean-centered age. Incorporating insights from EDA, three interactions terms are added in the model: `race` and `treat`, `educ_fac` and `race`, and `treat` and `agec`. Note that the variable `nodegree` (which comes from the original dataset) is dropped because to avoid the colinearity problem since it shares the same information as `educ_fac`.

The model assessment is done by plotting the residuals plots and the Q-Q Plot. The residual plots are roughly random and no discernible trend can be spotted. The Q-Q plot shows that the majority points follow the diagonal line well with some deviations at the beginning and the end. Therefore, it is reasonable to conclude that the linear model assumptions are not violated and the linear model is appropriate for the data.

However, as mentioned in EDA, there are potential outliers in the dataset. By assessing the leverage scores, influence scores, and standardized residuals, we find that indeed there are a few residuals that are outside

three standard deviations and have leverage scores greater than the threshold ($2(p+1)/n = 0.03$). However, none of the points have high influence. Since there are no influential point, it is reasonable to keep all observations and maintain the sample size.

The RMSE from 10-fold cross validation is 7610.53.

**Model Selection** To reduce model complexity and potentially reduce standard errors, the AIC forward model selection is performed on the initial model. The resulting model include `treat`, `agec`, `married`, and the interaction between `treat` and `agc`. F-test is then performed to test if the dropped variables from AIC stepwise selection have coefficients that are statistically different from zero collectively. The p-value from the test is large, therefore, we failed to reject the null hypothesis that these coefficients are zeros. They are, thereby, dropped from the initial model.

The resulting final model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

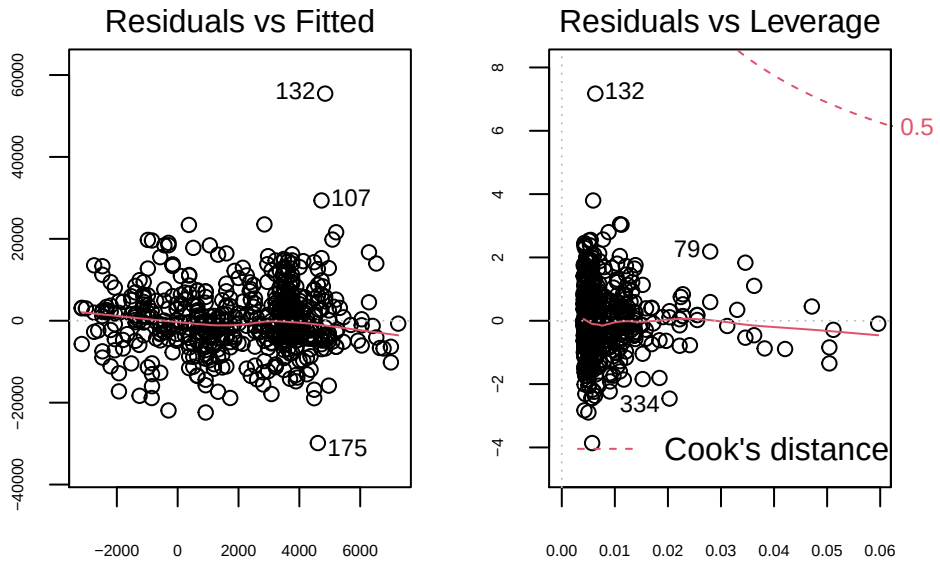
where $x_1$ is `treat`, $x_2$ mean-centered age, $x_3$ is `married`, and $x_4$ is the interaction between `treat` and mean-centered age.

**Final Model Results** The results from the final model is shown in the table below:

| | Estimate | s.e | p-value |
|---|---|---|---|
| Intercept | 2356.30 | 518.80 | 0 |
| factor(`treat`)1 | 2415.36 | 724.25 | 0.0009 |
| `agec` | -135.79 | 37.11 | 0.0002 |
| factor(`married`)1 | -1756.52 | 715.72 | 0.014 |
| `treat1:ageC` | 255.88 | 87.21 | 0.003 |
| Residual Standard Error | 7756 on 609 degrees of freedom | | |
| R-squared | 0.07 | | |

All variables are at least significant at 1% significance level. The intercept means that the expected wage increase from 1974 to 1978 is 2356.30 for an unmarried, 27-year-old person who is in the control group. All else the same, for one year increase in age, the wage increase from 1974 to 1878 is expected to decrease by 135.8 for someone in the control group. All else the same, the wage increase from 1974 to 1978 is expected to decrease by 1756.5 for a married person compared to an unmarried person. Furthermore, all else the same, compared to someone who is in the control group, receiving the "treatment" is associated with a wage increase of 2415 plus 255.88 times mean-centered age.

The final model assumptions are assessed by plotting residual plots and Q-Q plots. From the residual plots, the points are scattered roughly evenly around zero, and there is no discernible trend. Most points follow the diagonal line on the Q-Q plot with a few points deviate from the line at the beginning and the end. Therefore, it is reasonable to conclude that there is no obvious violation of linear model assumptions. Despite that a few observations are potentially outliers and have high leverage, no point has high influence. Therefore, no observation should be excluded for modeling.




Since VIF scores for all variables are low, it is safe to conclude that the model is unlikely to have multicolinearity problem.

Using 10-fold cross-validation, the model's RMSE is 7519.78.

**Conclusion**

Using the final model specified above, we find that receiving job training has a positive linear relationship with wage increase. For a 27-year-old person (the average age in the sample), receiving ob training is associated with 2415.36 increase in wage. We are 95% confident that all else the same, receiving job training is expected to increase wage by between 993 and 3837.7 for someone who has the average age in the sample. In addition, we also find that there is interaction effect between receiving treatment and age. In other words, the wage increase is expected to increase by an additional 255.88 times the mean-centered age. The 95% confidence interval for the interaction effect is between 84.6 and 427.2. We also find a significant negative relationship between marriage and wage increase. We are 95% confident that the expected wage increase decreases by between 350 and 3162 who someone who is married all else the same.

It is interesting to note that despite age has a negative relationship with wage increase, the positive interaction between age and treatment signifies that receiving treatment can reduce the negative effect of age. Furthermore, it is somewhat counter-intuitive to find that education does not seem to contribute to wage increase.

One of the problems with the model is the model fitness. The final model has low R-square at 0.07 and high residual standard error at 7756. The standard error for each coefficient is also large, particular for `treat` and `married`. The potential reason for this could be the high variance in wage increase between treatment and control groups, between married and unmarried people. For example, the standard deviation of wage increase for the treatment group is 8926, and the standard deviation of wage increase for married people is 7945.3. One solution for this is to increase the sample size and include more relevant variables for better goodness of fit.

# PART II

***Summary***

In this analysis, we build a logistic regression model to analyze whether receiving job training will increase the probability of workers receiving a positive wage. We do not find enough evidence to show that receiving job training affects the probability of having a positive wage. Besides that, We find that Blacks are more likely to receive a zero wage compared to other races. Specifically, the odds of receiving a positive wage in 1978 decreases by 48% compared to someone who is neither Black nor Hispanic.

***Introduction***

In the 1970s, researchers wanted to evaluate whether or not job training for disadvantaged workers had an effect on their wages. They collected data from workers who were randomly assigned either to receive training or not in the National Supported Work (NSW) Demonstration. For our analysis purpose, we use a subset of the original data. Throughout the analysis, we are motivated to find whether workers who received the job training tend to be more likely to receive a positive wage compared to those who did not receive training. Moreover, we would like to quantify the effect of job training on workers and find the range of the effects. Finally, we want to know if training effects differ by demographic groups and if there are other interesting relationships within data.

***Data***

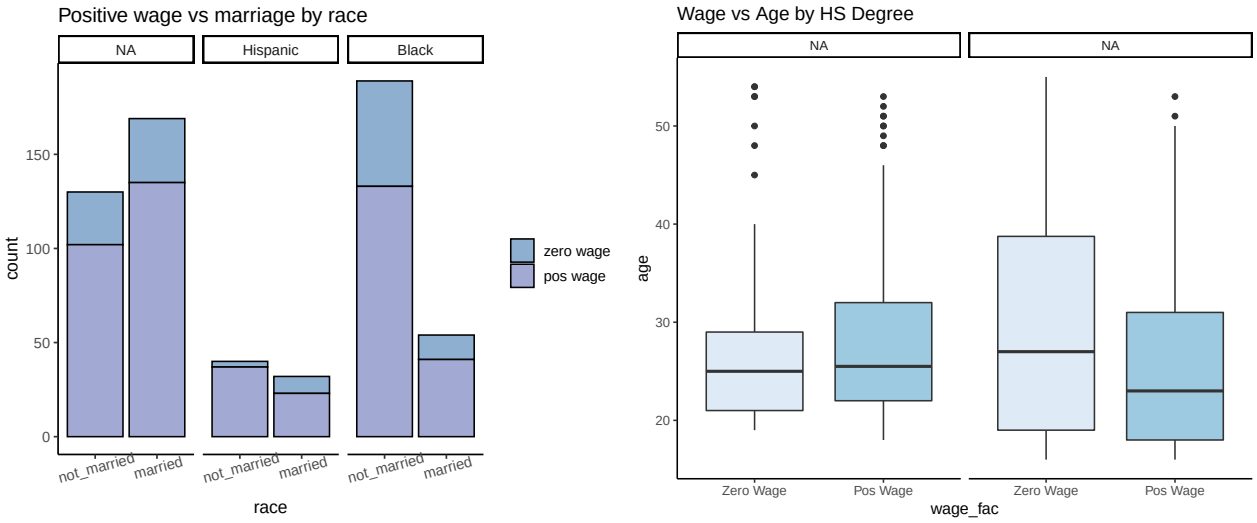
**Data Preperation & Summary** The dataset used in this analysis contains 614 observations on 7 variables. The response variable is `wage_fac`. It is a binary variable that takes the value "1" when an observation has a positive wage in 1978, and "0" otherwise. There are 143 observations with a zero wage, and 741 with a positive wage. The variable of interest is `treat`. It is also a binary variable that takes the value "1" when an observation is in the treatment group, and "0" in the control group. The `age` variable ranges from 16 to 55 years old, with an average of 27 years old. In this sample, the years of education received ranges from zero to 18 years with an average of 10. 227 observations are married (encoded "1") and 359 are not married (encoded "0"). More than half of the observations (387 out of 614) dropped out of high school (encoded as "1"). Finally, we constructed `race` variable by combining `black` and `hispan` variables. Therefore, `race` takes the value "0" for non-Hispanic and non-Black observations, "1" for Hispanic, and "2" for Black. We constructed the `race` variable because it can help us to explore the how demographics might interact with other variables to affect wage.

| | treat | age | educ | married | nodegree | wage_fac | race |
|---|---|---|---|---|---|---|---|
| | 0:429 | Min. :16.00 | Min. : 0.00 | 0:359 | Not drop-out:227 | Zero Wage:143 | Neither:299 |
| | 1:185 | 1st Qu.:20.00 | 1st Qu.: 9.00 | 1:255 | drop-out :387 | Pos Wage :471 | Hispan : 72 |
| | | Median :25.00 | Median :11.00 | | | | Black :243 |
| | | Mean :27.36 | Mean :10.27 | | | | |
| | | 3rd Qu.:32.00 | 3rd Qu.:12.00 | | | | |
| | | Max. :55.00 | Max. :18.00 | | | | |

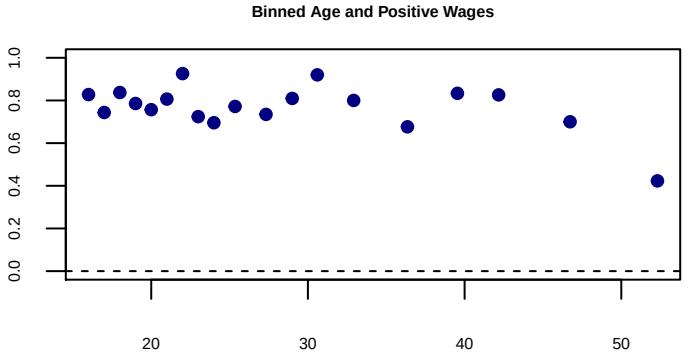
**Exploratory Data Analysis** In this sample, 75.7% of the people who are in the treatment group received positive wages in 1978 compared to 77.2% people in the control group who received positive wages. Although the difference is small, it implies little relationship between treatment and receiving positive wage. Out of all three races, Hispanics are the most likely to receive positive wages (83.3%), followed by non-Black and non_Hispanic (79.3%), and Blacks (71.6%). We also notice that Blacks are more likely to be in treatment group and Hispanics are more likely to be in the control group. Furthermore, we also find that that are potential interaction effects between `treat` and `race` because the trends differ between different races. However, it is important to point out that observations for Hispanic are limited. In fact, there is *zero* Hispanic

person in the treatment group who received a zero wage. Therefore, it is not reasonable to include an interaction between `race` and `treat`.

The plot "Positive wage vs marriage by race" shows that Blacks are less likely to be married compared to other races. Unmarried Blacks are more likely to have zero wages than other groups. This plot suggests that the relationship between receiving positive wages and marriage could differ by races. An interaction term between race and marriage may be necessary for model fitting. The plot "Wage vs Age by HS Degree" shows that for high school dropouts, younger people are more likely to receive positive. Although the age difference is small, the trend seems to be reversed for people who have high school degrees. This suggests there could potentially be interactions effects between `age` and `nodegree`.




Furthermore, the plot "Binned Positive Wages and Age" shows that binned positive wages shows a potential polynomial trend at around age 30 to 40. Therefore, it is useful to explore polynomial form of age in the model fitting.




### *Model*

**Initial Model** Incorporating insights from EDA, the full model is given as the following:

$$y_i \mid x_i \sim \ Bernoulli(\pi_i) \log(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{x_i}\boldsymbol{\beta},$$

where $y_i$ is `wage_fac`. It follows the Bernoulli Distribution that parameterized by $\pi_i$, which is the probability that a given observation has positive wage. $\boldsymbol{x_i}$ includes main effect variables: `treat`, `agec`, `agec_sq`, `educ`, `race`, `married`, and `nodegree`. `agec` and `agec_sq` are mean-centered age and squared mean-centered age. The squared `agec` is included to reflect the polynomial trend observed in the binned plot. In addition,

interaction term `agec` and `nodegree` are included to reflect our EDA findings that potentially `treat` and `married` could affect the probability of receive positive wage differently depending race.

The binned residual plots from the full model do not show any discernible trend. This suggests that no obvious violation of the model assumptions. Using 0.5 as the cut-off threshold for predicting `wage_fac` is "1", the model has an overall accuracy at around 0.783, with 0.998 sensitivity rate and 0.077 specificity rate. The model AUC is 0.626.

**Model Selection** To reduce model complexity, the AIC forward model selection is performed on the full model. The resulting model is given as:

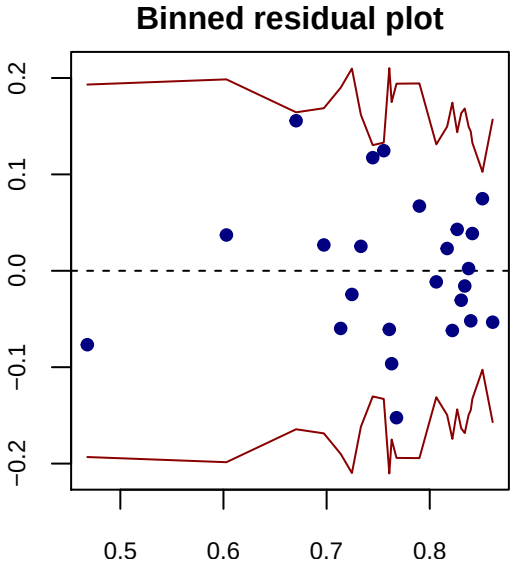
$$y_i \mid x_i \sim \ Bernoulli(\pi_i) \log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4},$$

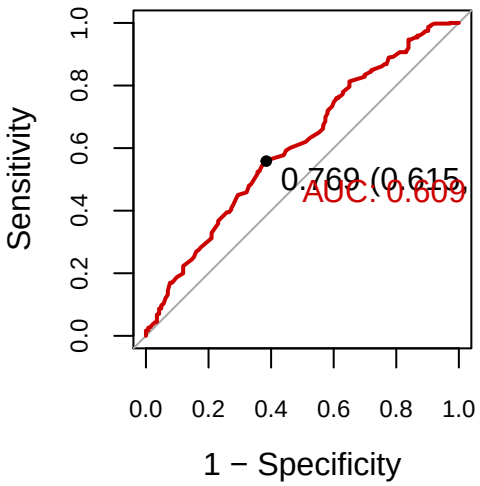
where $x_{i1}$ is mean-centered age (`agec`), $x_{i2}$ is squared mean-centered age (`agec_sq`), $x_{i3}$ indicates if an observation is in the treatment group (`treat`), $x_{i4}$ indicates whether someone is Hispanic, Black, or Neither (`race`).

F-test is then performed to test if the dropped variables from the AIC forward selection have coefficients that are statistically different from zero collectively. The p-value from the test is larger than 0.20, so we failed to reject the null hypothesis. We conclude that the drop variables are not statistically different from zero, and it is reasonable to drop them.

Binned residual plots are used to assess model assumptions. The residual plots are random and no obvious trend can be observed. Therefore, it is reasonable to assume the model is appropriate. Furthermore, the VIF for all predictors range from 1.1 to 2.4. The VIF values are relatively small, so we conclude that the model is unlikely to have have multicolinearity problem.

In terms of the model validation, the model achieves 0.783 overall accuracy using 0.5 as the cut-off threshold for predicting `wage_fac` equal to "1". Using the same threshold, the model achieves 0.998 sensitivity and 0.08 specificity. Using ROC, the ideal cut-off threshold is 0.769 which will achieve 0.615 specificity and 0.558 sensitivity. Overall, the model has AUC at 0.609.



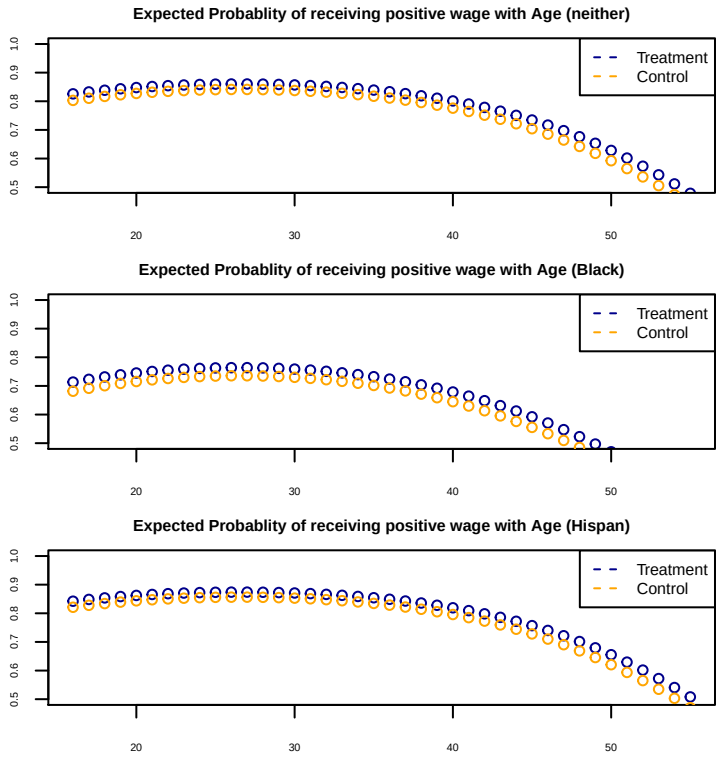

**Final Model & Results** The results of the final model is shown below.

| | Estimate | Odds Ratio | Standard Error | p-value |
|---|---|---|---|---|
| *Intercept* | 1.666 | 5.29 | 0.186 | 0 |

| | Estimate | Odds Ratio | Standard Error | p-value |
|---|---|---|---|---|
| `agec` | -0.0038 | 0.996 | 0.0138 | 0.783 |
| `agec_sq` | -0.0024 | 0.998 | 0.0009 | 0.012 |
| `treat1` | 0.151 | 1.163 | 0.267 | 0.57 |
| `race`Hispan | 0.118 | 1.125 | 0.354 | 0.74 |
| `race`Black | -0.645 | 0.525 | 0.2555 | 0.01 |

The intercept is interpreted as the expected odds of receiving positive wage for someone who has the same age as the sample (~27 years old), in the control group, and is neither Hispanic nor Black is 5.29. For the coefficient for `treat`, it means that holding all else the same, the odds of receiving positive wage for someone who is the treatment group increases by around 16% compared to that of the control group. However, the coefficient is not statistically significant. Despite not significant, the interpretation for Hispanic is that holding all else the same, the odds of receiving positive wage increases by 12% for Hispanic, compared to that of someone who is neither Hispanic nor Black. The coefficient for Black means that all else the same, the odds of receiving positive wage for Black is multiplied by 0.52 compared to that of someone who is neither Hispanic nor Black. Note that this coefficient is significant at 0.05 significance level. Finally, the interpretation for the age-related variables is somewhat involved because of the quadratic term. The plot below helps to visualize the effects of age.

In the graphs below plot the expected probability of receiving positive wage for people age from 16 to 55 years old. The plots show a clear concave relationship between predicted probability and age for all races. It is also interesting to note that although the probability of receiving positive wage is slightly higher for the treatment group, the improvement is small. This corresponds to the regression results that `treat` is not statistically significant. In addition, the predicted probabilities for Black are smaller than the other two groups regardless age. This also corresponds to the regression results that the odds of receiving positive wage is significantly lower than that of the neither race group.



***Conclusion***

Overall, this analysis does not find enough evidence to conclude that there is a statistically significant relationship between the probability of receiving positive wage and additional job training. Although not statistically significant, the point estimation for the odds ratio between treatment and control group shows a slight positive relationship between receiving training and the probability of receiving positive wage. It is estimated that the odds of receiving positive wage is multiplied by between 0.69 and 1.96 with 95% confidence level. We also find that, in general, Blacks have lower odds of receiving positive wage than other races. Specifically, the odds of receiving positive wage for Black is multiplied by between 0.32 and 0.87 compared with someone who is neither Black nor Hispanic with 95% confidence level.

It is interesting to point out that the odds of receiving positive wage has a significant negative quadratic relationship with age. This implies that the odds of having positive wage tend to increase with age only up to a certain threshold, then the odds tend to decrease as age increases.

It is important to note that one of the limitations with the model is the lack of data for certain groups. For example, there are only 72 Hispanic in the sample. Out of the 72 Hispanic observations, there are only 11 people who were in the treatment group. Furthermore, there are zero person in this group who had zero wage. The lack of data makes it impossible for us to explore the potential interaction effects between treatment and race. In addition, the model has low specificity and AUC even with the full model. The AUC for the final model is 0.609 and the AUC for the full model is 0.626 These indicates the model fitness is not ideal. This could be resulted from the fact that the majority of the sample has positive wage (76.7%). Because of the bias in the sample, the model tends to predict positive wage. To improve the model, one solution is to expand the sample size and obtain more observations on diverse ranges of people, such as more Hispanic and more zero wage observations.

## R Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(car)
library(arm)
library(pROC)
library(e1071)
library(caret)
library(knitr)
library(dplyr)
library(ggplot2)
library(pander)
library(gridExtra)
library(kableExtra)
library(stargazer)
opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)

# Data import -----------
df = read.csv("~/Desktop/DukeFA20/IDS702/Team Assignment1/lalondedata.txt",
    header = TRUE, sep = ",")

# Data preparation -----------------
df$treat <- as.factor(df$treat)
df$black <- as.factor(df$black)
df$hispan <- as.factor(df$hispan)
df$married <- as.factor(df$married)
df$nodegree <- as.factor(df$nodegree)
df$nodegree <- factor(df$nodegree, levels = c(0, 1), labels = c("Not drop-out",
    "drop-out"))

# Response Variable for MLR (PART I)
df$wage_inc <- with(df, re78 - re74)

# Response variable for logistic regression (PART II)
df$wage <- with(df, ifelse(re78 > 0, 1, 0))
df$wage_fac <- factor(df$wage, levels = c(0, 1), labels = c("Zero Wage",
    "Pos Wage"))

# Collapse black, hispan -> race table(df$black,df$hispan)
df$race = with(df, ifelse(black == 0 & hispan == 0, 0, ifelse(black ==
    0 & hispan == 1, 1, 2)))
df$race <- factor(df$race, levels = c(0, 1, 2), labels = c("Neither",
    "Hispan", "Black"))

# Centering age
df$ageC <- df$age - mean(df$age)
df$ageC_sq <- df$ageC^2

# collapse educ
df$educ_fac <- with(df, ifelse(educ < 9, 0, ifelse(educ < 13,
    1, 2)))
df$educ_fac <- factor(df$educ_fac, levels = c(0, 1, 2), labels = c("Middle School",
    "High School", "College"))
```

```r
# educ and wage_inc

# ggplot(df, aes(wage_inc, group = educ_fac)) +
# geom_density()


kable(summary(df[, colnames(df) %in% c("wage_inc", "educ_fac",
    "age", "race", "married", "treat")]))


p1 = ggplot(df, aes(wage_inc)) + geom_histogram(aes(y = ..density..),
    color = "black") + geom_density(alpha = 0.2, fill = "blue") +
    labs(title = "Distribution of wage increase") + theme_classic(base_size = 5)


# educ_fac v wage by race

p2 = ggplot(df, aes(x = educ_fac, y = wage_inc, fill = race)) +
    geom_boxplot() + scale_fill_brewer(palette = "Blues") + labs(title = "Wage increase vs educ by race
    x = "Eduction", y = "wage increaase") + theme_classic(base_size = 8) +
    theme(legend.position = "none", axis.text.x = element_text(angle = 15)) +
    facet_wrap(~race)

# wage_inc vs age by treat
p3 = ggplot(df, aes(x = ageC, y = wage_inc)) + geom_point(alpha = 0.5,
    colour = "blue4") + scale_fill_brewer(palette = "Blues") +
    geom_smooth(method = "lm", col = "red3") + labs(title = "Wage vs age by treat",
    x = "mean-centered age", y = "wage increase") + theme_classic(base_size = 8) +
    theme(legend.position = "none") + facet_wrap(~treat, labeller = as_labeller(c(`0` = "Control Group"
    `1` = "Treatment Group")))

# wage_inc vs race by treat
p4 = ggplot(df, aes(x = race, y = wage_inc, fill = race)) + geom_boxplot() +
    scale_fill_brewer(palette = "Blues") + labs(title = "Wage increase vs race by treat",
    x = "race", y = "wage increase") + theme_classic(base_size = 8) +
    theme(legend.position = "none") + facet_wrap(~treat, labeller = as_labeller(c(`0` = "Control Group"
    `1` = "Treatment Group")))


grid.arrange(p1, p2, p3, p4, ncol = 2)


model_4 <- lm(data = df, wage_inc ~ ageC + educ_fac + treat +
    race + married + race * treat + educ_fac * race + treat *
    ageC)
# summary(model_4)

set.seed(123)

K = 10
dfr <- df[sample(nrow(df)), ]
RMSE <- matrix(0, nrow = K, ncol = 1)
```

```r
kth_fold <- cut(seq(1, nrow(dfr)), breaks = K, labels = FALSE)

for (k in 1:K) {
    test_index = which(kth_fold == k)
    train = dfr[-test_index, ]
    test = dfr[test_index, ]
    fit = lm(data = dfr, wage_inc ~ ageC + educ_fac + treat +
        race + married + race * treat + educ_fac * race + treat *
        ageC)
    predictions = predict(fit, test)
    RMSE[k, ] = sqrt(mean((predictions - test$wage_inc)^2))
}

k_rmse = mean(RMSE)


null_model <- lm(wage_inc ~ treat, data = df)
model_stepwise <- step(null_model, scope = formula(model_4),
    direction = "forward", trace = 0)


summary(model_stepwise)
final_form <- wage_inc ~ treat + ageC + married + treat:ageC
final_model <- lm(final_form, data = df)
# summary(final_model)

# anova(model_4, final_model)



# summary(final_model)

par(mfrow = c(1, 2), mar = c(1.8, 1.8, 1.8, 1.8))

plot(final_model, which = 1, cex.main = 0.5, cex.lab = 0.5, cex.axis = 0.5)
plot(final_model, which = 5, cex.main = 0.5, cex.lab = 0.5, cex.axis = 0.5)

# plot(final_model)
vif(final_model)

K = 10
dfr <- df[sample(nrow(df)), ]
RMSE <- matrix(0, nrow = K, ncol = 1)

kth_fold <- cut(seq(1, nrow(dfr)), breaks = K, labels = FALSE)

for (k in 1:K) {
    test_index = which(kth_fold == k)
    train = dfr[-test_index, ]
    test = dfr[test_index, ]
    final_form <- wage_inc ~ treat + ageC + married + treat:ageC
    final_model <- lm(final_form, data = df)
    predictions = predict(fit, test)
```

```
    RMSE[k, ] = sqrt(mean((predictions - test$wage_inc)^2))
}

k_rmse = mean(RMSE)


columns = c("wage_fac", "treat", "age", "race", "married", "nodegree",
    "educ")

sumdf = df[, colnames(df) %in% columns]

opts <- options(knitr.kable.NA = "")

knitr::kable(summary(sumdf))

### wage_fac vs age by nodegree
p1 = ggplot(df, aes(x = wage_fac, y = age, fill = wage_fac)) +
    geom_boxplot() + scale_fill_brewer(palette = "Blues") + labs(title = "Wage vs Age by HS Degree") +
    theme_classic() + facet_wrap(~nodegree, labeller = as_labeller(c(`0` = "HS Degree",
    `1` = "HS Dropout"))) + theme(legend.position = "none")


### wage_fac vs married by race ------interesting proportion
### for hispan
p2 = ggplot(df, aes(x = married, fill = wage_fac)) + geom_bar(color = "black") +
    scale_x_discrete(labels = c("not_married", "married")) +
    scale_fill_discrete(labels = c("zero wage", "pos wage"),
        h = c(240, 260), c = 35, l = 70) + # scale_fill_brewer(palette='Blues') +
labs(title = "Positive wage vs marriage by race", x = "race") +
    theme_classic(base_size = 12) + theme(legend.title = element_blank()) +
    facet_wrap(~race, labeller = as_labeller(c(`Non Black and Hispan` = "Neither",
        Hispan = "Hispanic", Black = "Black"))) + theme(axis.text.x = element_text(angle = 15))

grid.arrange(p2, p1, nrow = 1)


par(mar = c(2, 2, 2, 2))
# binned plot of age -- polynomial terms

binnedplot(y = df$wage, df$age, xlab = "age", ylim = c(0, 1),
    col.pts = "navy", ylab = "Wage greater than 0?", main = "Binned Age and Positive Wages",
    col.int = "white", cex.main = 0.5, cex.lab = 0.5, cex.axis = 0.5)


df$age_sq = df$age^2
df$age_tri = df$age^3
df$agec = df$age - mean(df$age)
df$agec_sq = (df$agec)^2

form = wage ~ agec + agec_sq + educ + nodegree + treat + married +
    race + agec * nodegree

model <- glm(data = df, form, family = binomial)
```

```r
# summary(model)

rawresid = residuals(model, "resp")

# Model Assessment
# binnedplot(x=fitted(model),y=rawresid,xlab='Pred.
# probabilities', col.int='red4',ylab='Avg.
# residuals',main='Binned residual plot',col.pts='navy')
# binnedplot(x=df$age,y=rawresid,xlab='age',
# col.int='red4',ylab='Avg. residuals',main='Binned residual
# plot',col.pts='navy')
# binnedplot(x=df$educ,y=rawresid,xlab='educ',
# col.int='red4',ylab='Avg. residuals',main='Binned residual
# plot',col.pts='navy')

# Confusion matrix
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(model) >=
    0.5, "1", "0")), as.factor(df$wage), positive = "1")

# Conf_mat$table Conf_mat$overall['Accuracy'];
# Conf_mat$byClass[c('Sensitivity','Specificity')] #True
# positive rate and True negative rate

# ROC
# roc(df$wage,fitted(model),plot=T,print.thres='best',legacy.axes=T,
# print.auc =T,col='red3')


# AIC ++++++++ Final Model
null_model <- glm(wage ~ treat, data = df, family = binomial)
aic_model <- step(null_model, scope = formula(model), direction = "forward",
    trace = 0)


# summary(aic_model)

final_form = wage ~ agec + agec_sq + treat + race
final_model = glm(final_form, df, family = binomial)
# summary(final_model)


# Model Assessment anova(model, final_model, test = 'Chisq')
# vif(final_model)


# Model Assessment
# binnedplot(x=fitted(final_model),y=rawresid,xlab='Pred.
# probabilities', col.int='red4',ylab='Avg.
# residuals',main='Binned residual plot',col.pts='navy')
# binnedplot(x=df$age,y=rawresid,xlab='age',
# col.int='red4',ylab='Avg. residuals',main='Binned residual
# plot',col.pts='navy')
# binnedplot(x=df$educ,y=rawresid,xlab='educ',
```

```r
# col.int='red4',ylab='Avg. residuals',main='Binned residual
# plot',col.pts='navy')

# Model Validation

# Confusion matrix
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(final_model) >=
    0.5, "1", "0")), as.factor(df$wage), positive = "1")

Conf_mat$table
Conf_mat$overall["Accuracy"]
Conf_mat$byClass[c("Sensitivity", "Specificity")]  #True positive rate and True negative rate

# ROC
# roc(df$wage,fitted(final_model),plot=T,print.thres='best',legacy.axes=T,
# print.auc =T,col='red3') Plots to show

par(mfrow = c(1, 2), mar = c(1.8, 1.8, 1.8, 1.8))
rawresid = residuals(final_model, "resp")
binnedplot(x = fitted(final_model), y = rawresid, xlab = "Pred. probabilities",
    col.int = "red4", ylab = "Avg. residuals", main = "Binned residual plot",
    col.pts = "navy", cex.axis = 0.75, cex.main = 1)

roc(df$wage, fitted(final_model), plot = T, print.thres = "best",
    legacy.axes = T, print.auc = T, col = "red3", cex.axis = 0.75)

# library(sjPlot) library(sjmisc) library(sjlabelled)
# tab_model(final_model)

library(stargazer)

# stargazer(final_model, apply.coef = exp, apply.se = exp,
# apply.ci = exp, title='Regression Results', # align=TRUE,
# single.row=TRUE, header = FALSE)

par(mfrow = c(3, 1), mar = c(2, 2, 2, 2), mgp = c(3, 1, 0))
newage <- seq(from = 16, to = 55)
newagec <- newage - mean(df$age)
newage_sq <- newagec^2
newdata <- data.frame(matrix(0, nrow = length(newagec), ncol = 4))
names(newdata) <- c("agec", "agec_sq", "treat", "race")
newdata$agec <- newagec
newdata$agec_sq <- newage_sq

newdata$treat <- 1
newdata$treat = as.factor(newdata$treat)
newdata$race = "Neither"
newdata$race = as.factor(newdata$race)
# Since we use mean-centered predictors, the rows in the new
# dataset correspond to people with average values of
# seniority, age, and education.
preds_treat <- predict(final_model, newdata, "resp")
newdata$treat <- 0
```

```r
newdata$treat = as.factor(newdata$treat)
preds_control <- predict(final_model, newdata, "resp")

plot(y = preds_treat, x = newage, xlab = "age", ylab = "Predicted Wages",
    main = "Expected Probablity of receiving positive wage with Age (neither)",
    col = "darkblue", ylim = c(0.5, 1), cex.main = 0.7, cex.lab = 0.5,
    cex.axis = 0.5)
points(y = preds_control, x = newage, col = "orange")
legend("topright", c("Treatment", "Control"), col = c("darkblue",
    "orange"), lty = c(2, 2), cex = 0.75)

# Black
newdata$treat <- 1
newdata$treat = as.factor(newdata$treat)
newdata$race = "Black"
newdata$race = as.factor(newdata$race)
preds_treat <- predict(final_model, newdata, "resp")
newdata$treat <- 0
newdata$treat = as.factor(newdata$treat)
preds_control <- predict(final_model, newdata, "resp")
plot(y = preds_treat, x = newage, xlab = "age", ylab = "Predicted Wages",
    main = "Expected Probablity of receiving positive wage with Age (Black)",
    col = "darkblue", ylim = c(0.5, 1), cex.main = 0.7, cex.lab = 0.5,
    cex.axis = 0.5)
points(y = preds_control, x = newage, col = "orange")
legend("topright", c("Treatment", "Control"), col = c("darkblue",
    "orange"), lty = c(2, 2), cex = 0.75)

# Hispan
newdata$treat <- 1
newdata$treat = as.factor(newdata$treat)
newdata$race = "Hispan"
newdata$race = as.factor(newdata$race)
preds_treat <- predict(final_model, newdata, "resp")
newdata$treat <- 0
newdata$treat = as.factor(newdata$treat)
preds_control <- predict(final_model, newdata, "resp")
plot(y = preds_treat, x = newage, xlab = "age", ylab = "Predicted Wages",
    main = "Expected Probablity of receiving positive wage with Age (Hispan)",
    col = "darkblue", ylim = c(0.5, 1), cex.main = 0.7, cex.lab = 0.5,
    cex.axis = 0.5)
points(y = preds_control, x = newage, col = "orange")
legend("topright", c("Treatment", "Control"), col = c("darkblue",
    "orange"), lty = c(2, 2), cex = 0.75)
```