Effects of Job Training on Wages

Akshay Punwatkar, Melody(Xinwen) Li, Derek Wales, Andrew Patterson, Tzu-Chun (Angela) Hsieh

10/4/2019

Part I: Analysis of the effects of Training on Wages in 1978

SUMMARY

For this assignment we were working in our MIDS Assigned Teams, looking at a Dataset to determine whether vocational training through the National Support Work (NSW) helps people to gain higher wage. This project was carried out in order to answer several key research questions on this work training's effectiveness. To do this we used a series of methods, including Teamwork, Critical Thinking, Exploratory Data Analysis, and Logistic Regression which ultimately lead to the model below.

INTRODUCTION

This project seeks to answer the question about whether job training helps people to earn higher wages compared to a group of people who were unemployed in 1976 whose income in 1975 was below the poverty level. A linear regression was constructed to model the actual wage increase using various predictor variables, including treatment.

DATA

Data used in the analysis containing **614 observations** was a **non-randomized** subset of the original study. Also, observations in the data constitue **male** workers only. **Wages from 1975 were NOT included** in the study because they carried certain ambiguity since they overlapped the training period (1975-77).

Data Dictionary :

	Description
treat	Indicator whether or not a worker went through training
age	Age of the Worker in years.
educ	Years of education
black	Indicator whether the worker belong to Black race or not
hisp	Indicator whether the worker belong to Hispanic race or not
married	Indicator whether the worker was married or not
nodegree	Indicator whether the worker dropped out of high school or not
re74	Worker's real annual earnings in 1974.
re75	Worker's real annual earnings in 1975.
re78	Worker's real annual earnings in 1978.

Data Processing :

Based on the intial analysis, the wages in 1978 **did not appear to follow the normal distribution** leading to right skewness, which can be attributed to almost 1/5 of wages being zero. A decision was made to remove the zero value wages from 1978. The resulting distribution showed some improvement towards being a normal distribution.

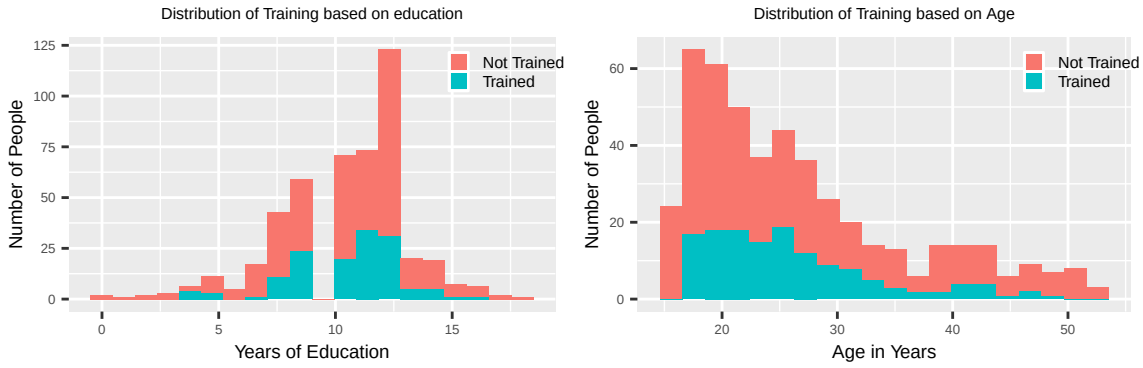
Data Transformation :

None of the data was transformed for the analysis. Log tranformation of the reponse variable (wages in 1978) was analysed and appeared to have a left skewness (as shown in the figure above), so based on this a decision was made to use the response variables without any transformations.

Exploratory data analysis :

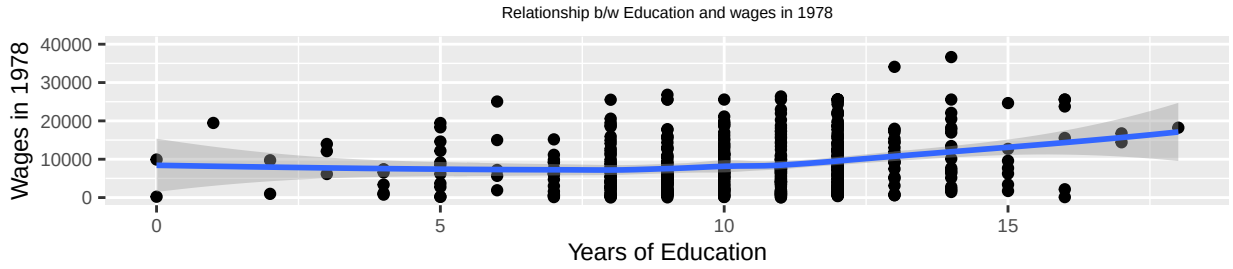
Based on the Initial data analysis, the following was concluded :

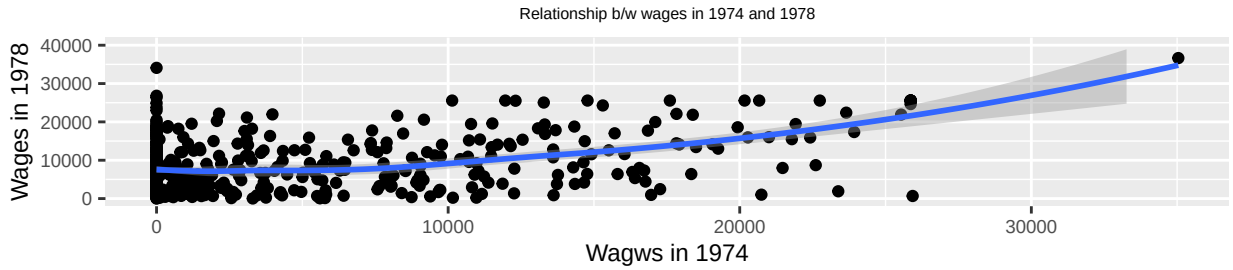
- It was evident that people within a certain **education range (7-12 years)** and people below a certain **age (<30 years)** were more involved in the experiment. Also, both **age** and **education** appeared to have a nearly **constant linear** relationship with wages in 1978.
- Although, **wages in 1974** did appear to have an **incrementing linear relationship** with wages in 1978.
- **More** people with **black** race **received training** as compared to non-black.
- **Very few** people with **Hispanic** race **received training** as compared to non-hispanic.
- More non-married people received training as compared to married ones.
- More highschool drop-out people received training as compared to others.
- None of the continous variables appeared to be highly correlated.



Association with and within the Categorical Variables :

Var_1	Var_2	p.val	Significant	Var___1	Var___2	p.val_	Significant_
re78	treat	0.34459	No	treat	black	1.466012052378e-49	Yes
re78	black	0.40540	Yes	treat	hispan	0.00532282290801077	Yes
re78	hispan	0.12165	No	treat	married	1.612783524077e-13	Yes
re78	married	0.39122	No	treat	nodegree	0.0113435849379384	Yes
re78	nodegree	0.38903	No				





MODELING

- The initial full model was performed using all the available variables (execpt wages in 1975) which resulted in a *adjusted R-Square* value of **15.04%** with **age**, **educ** and **re74** as significant. A stepwise-AIC method was applied over the full model and the resulting model included treatment, education, black, and wages in 1974 (re74).
- A model was then constructed with the 4 variables based on stepwise model construction, resulting in a model with an *adjusted R-Square* of **14%**. However, **Black** did NOT appear to be significant, and this result was **confirmed in an F-test** when compared to another model with the 3 covariates (excluding black race).
- Based on initial analysis, which showed some kind of relationship between **wages in 1978** and **age**, **age** was added back into the model. The significance of age in the model was confirmed by an F-Test.
- For the improvement of the model performace, several interactions between covariates were tested but none of the interactions appeared to be statistically significant.
- Based on the evaluation of different models with different combinations of covariates and interactions, the **Final model** was constructed using **treat**, **age**, **educ** and **re74**, resulting in an *adjusted R-Square* for **16.25%**. Residual plots confirmed this model did not violate normality and independence. Equal variance and linearity do, however, appear to be violated. This could be because of a lack of data, especially since all zero re78 data points were removed. For these reasons, these violations were considered acceptable for the final model.

Final Model :

$$\hat{Re}_{78} = \hat{\beta}_0 + \hat{\beta}_1 Treat_i + \hat{\beta}_2 Educ_i + \hat{\beta}_3 Re74_i + \hat{\beta}_4 Age_i$$

	estimate	std.error	statistic	p.value	2.5 %	97.5 %
(Intercept)	294.2086004	1706.5336199	0.1724013	0.8631969	-3059.245526	3647.6627268
treatTrained	730.8885156	713.2212346	1.0247711	0.3060029	-670.639508	2132.4165396
educ	395.5808344	125.5721876	3.1502265	0.0017364	148.822982	642.3386865
re74	0.3308532	0.0536771	6.1637647	0.0000000	0.225374	0.4363323
age	97.4638979	38.0796967	2.5594715	0.0107977	22.634715	172.2930806

RESULTS

In the final model,

- **treatment** did NOT appear to be a significant covariate for predicting wages in 1978. However, since the primary objective of the analysis was to quantify the effects of training, **treat** was included in the final model. Quantitively, given other variables are controlled, those who received treatment received **730.89** more dollars in 1978 than workers who did not receive training. The 95% confidence interval for the effect of treatment on wages was -670.00 dollars for the lower bound and 2128.78 dollars for the upper bound. This could show the inconsistency of the effects of training i.e. with a 95% confidence we can conclude that training might have a negative impact on wages (as low as -$670) or a positive impact on the wage (as high as $2100).
- As for **demographics**, none of the racial groups turned out to be significant; this indicates that race does not have an effect on actual wages in 1978.

- But for other **demographics**, **age** does appear to be a significant variable which was highlighted during the initial analysis. With every unit increase in age, the predicted pay in 1978 increased by 97.46 dollars according to the final model, although with 95% confidence we can say that this change could be as low as $22 or as high as $172.

- **Education** also appeared significant for the prediction of wages in 1978; with every unit increase in the number of years of education, predicted wages increases by 395.58 dollars with a 95% confidence interval of ($148 - $642)

- Finally, **workers wages in 74** also appeared to be significant and can be associated with higher wages in 1978; with every unit increase in the wage of 1974, predicted wage increases by 33 cents with a 95% confidence interval of (22-43 cents).

CONCLUSION

From this model, the job training treatment in this study is associated with an increase in pay as compared to those who did not receive treatment.

Limitations of this model include a lack of original data- removing those who did not make any money in 1978 reduced the data by about 140 samples. Related to this, this model does not determine if this job training actually helps someone to find a job. A future project could be to analyze this data and determine if the job training actually helps someone to get a job.

From this, interesting future projects could be to compare different job training programs with each other to help determine the most effective way to incerease oppurtunity and wages for people who receive the training.

Part II: Effect of Training on Employment Status

Introduction 2

This project seeks to answer the question about whether job training results in an increased odds of being employed compared to a group who did not receive job training. All people tested had an income below the poverty line in 1975. A logistic regression model was constructed to model the effect of predictor variables on the odds of employment status.

Data 2

Data was received as a .txt filed called lalondedata.txt. This data was imported into R, treating the data as a .csv file. For this data, a new column called emp78 was created based on the re78 column (wage data from 1978). For this new column, if the 1978 data was above zero, then that entry was assigned a value of 1 in the new column; if the 1978 wage data was 0, that entry was assigned a value of 0 in emp78. Age and income in 1974 (re74) were also mean centered to facilitate interpretation of the model. No other changes were made to the data.

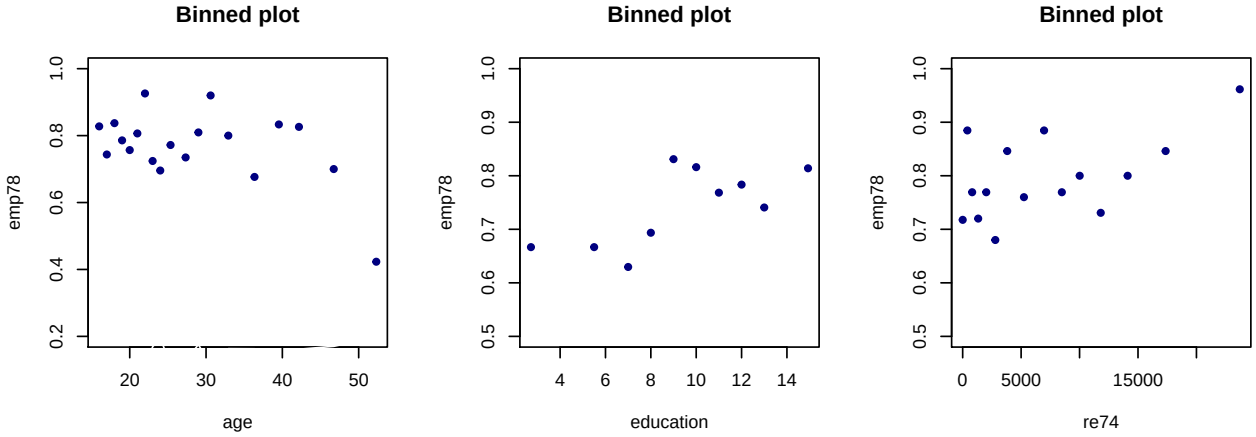
Explanatory Data Analysis

Chi-squared was performed between emp78 as compared to five factor variables to check for multicolinearity. The five variables were treatment, black, Hispanic, married, and no degree. Of these, only black and emp78 have a significant p-value indicating they are dependent in relation to each other. Another Chi-squared test was performed on treatment and the other four categorical variables: black, Hispanic, married, and no degree. All contain significant p values, indicating they are all dependent on each other. Therefore, we tested these interaction terms while constructing the model. Below is the table of the results of the Chi-squared tests:

Variable1A	Variable2A	p.valueA	SignificantA	Variable1B	Variable2B	p.valueB	SignificantB
emp78	treat	0.7686215	No	treat	black	1.466012052378e-49	Yes
emp78	black	0.0200977	Yes	treat	hispan	0.00532282290801077	Yes
emp78	hispan	0.2052281	No	treat	married	1.612783524077e-13	Yes
emp78	married	0.5756078	No	treat	nodegree	0.0113435849379384	Yes
emp78	nodegree	0.5052997	No				

Boxplots were constructed for emp78 on the quantitive variables (age, education, and re74). These box plots were also split between treatment and non-treatment plots. From these box plots, there is evidence of an interaction between treatment and age. This interaction was also considered when constructing the model.

Binnedplots for the continuous variables age, educ, and re74 were also constructed. From these plots, age does not follow the logit function pattern; there is a wave trend in the plot. Both education and re74 do appear to follow the logit pattern. From these observations of age, alternative strategies of handling this covariate were considered in model building, discussed below.

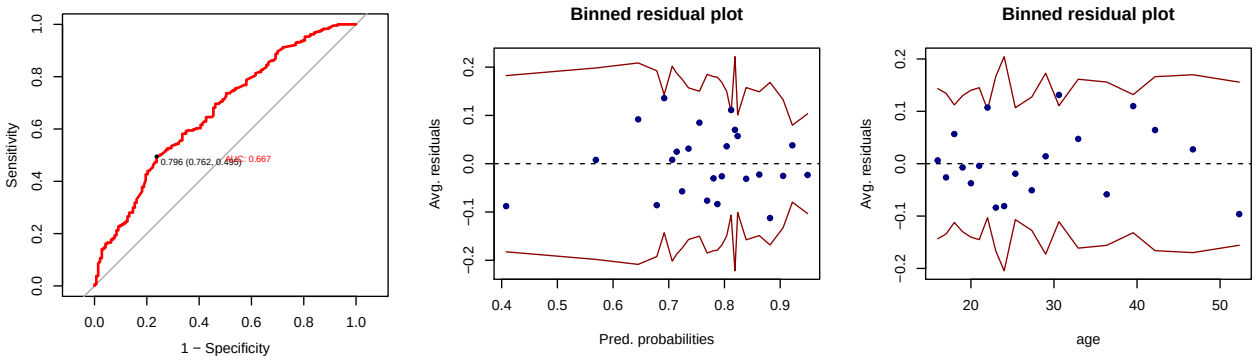


Models

Quantitive covariates age and income in 1974 (re74) were centered for easier interpretation in these models. The full model included age, education, treatment, whether the person was black, whether the person was Hispanic, whether the person was married, whether the person did not have a degree, wage data from 1974 (re74), and interactions between all of these covariates. This was used to predict the log odds of employment in 1978. A step AIC model construction generated a model using age, re74, black, treatment, the interaction between black and treatment, the interaction between re74 and treatment, and the interaction between age and treatment as covariates.

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```



The binned plot of residual vs predicted probability seems acceptable without any trends. However, as discussed in the EDA, the previous pattern of age still exists in the model. This means that the model loses information on age. To try and correct for this, age was split into two groups, one below 30 years old and one above 30 years old. This converted the continuous age variable into a two factor variable. Testing the model with the new age predictor did not improve the model, and age was not significant. Therefore, age as a continuous variable was used in the models.

Model Selection

The full model had an accuracy of 0.634, which is higher than the step constructed model accuracy of 0.611. This indicates the full model is actually slightly more accurate than the constructed model but with an offset of containing much more predictor variables. For the full model, sensitivity was 0.611, compared to the sensitivity of 0.616 for the constructed model. Both sensitivities are reasonably high. The specificity of the full model was 0.706, and the specificity of the constructed model was 0.611. The full model has a higher specificity, indicating the full model has less false negatives than the constructed model. For the full model, AUC was 0.730, while the AUC for the 0.667. Despite the lower values of the constructed model compared to the full model, the constructed model was ultimately used as the final model. This was because the values for the constructed model were acceptable while removing many insignficant interactions and covariates.

Additionally, Chi-squared tests were performed on the final model comparing the final model with a single removed interaction term (two remaining). This was done for every interaction term, and then for all the remaining predictors. All p-values were significant, indicating all of the interactions are significant in the model.

Equation for the final model:

$$log(\frac{Employed}{1-Employed}) = B_0 + B_1 Treatment_{yes} + B_2 Age + B_3 Black + B_4 Re_{74} + B_5 Re_{74} * Black + B_6 Age * Treatment_{yes} + B_7 Black * Treatment_{yes}$$

Table describing the final model: % latex table generated in R 3.6.0 by xtable 1.8-4 package % Fri Oct 4 20:05:45 2019

	term	estimate	std.error	statistic	p.value	conf.low	conf.high
1	(Intercept)	3.84570	0.14148	9.52055	0.00000	2.93518	5.11606
2	age_c	0.94582	0.01118	-4.98049	0.00000	0.92497	0.96656
3	re74_c	1.00010	0.00002	4.00768	0.00006	1.00005	1.00015
4	black1	0.61709	0.28439	-1.69751	0.08960	0.35554	1.08822
5	treat1	5.09007	0.75574	2.15323	0.03130	1.43659	32.46978
6	re74_c:black1	0.99992	0.00004	-2.03272	0.04208	0.99985	1.00000
7	age_c:treat1	1.08062	0.02754	2.81569	0.00487	1.02537	1.14304
8	black1:treat1	0.23831	0.81263	-1.76487	0.07759	0.03473	0.97510

Results and Interpretations

All the coefficient estimates in the table were exponentiated.

The intercept of 3.846 has a significant p-value. This intercept means that a male worker with average age, average '74 salary, non-black race, and does not receive treatment would have odds of employment in 78 at 3.846. For the confidence interval, there is a 95% confidence that the intercept will lie between 2.935 and 5.116.

The coefficient of centered '74 salary is 1.0001 with a significant p-value, which means that holding everything else constant, one year increase in the average '74 salary would have 100.01% increase of the odds of employment in 1978.

The coefficient of receving training is 5.09 with a significant p-value of 0.03, which means that holding everything else constant, receving training would have 500% increase of the odds of been employed in 1978 compare to someone does not receve training and with 95% confidence that the odds will lie between 1.43659 and 32.46978. The range of the effect of training on the odds of employment is very broad. This means that training can have a positive effect on employment status.

The coefficient of the interaction term (age * treatment) is 1.08 with a significant p-value of 0.004, which means that holding everything else constant, one year increase in the age with training been 1 would have a 8% increase in the odds of been employed in 1978 compare to one year increase in age without receving training.

For the demographic groups, the coefficient of centered age is 0.946 with a significant p-value, and the coefficient means that holding everything else constant, a one year increase in age would have 94.6% increase of the odds of employment in 1978 and 95% confidence that the odds will lie between 0.925 and 0.967. The interpretation of this result is that as age increases, odds of employment increases.

For demographic groups, while black is a covariate in the final model, it is not considered a signficant covariate due to a higher p-value than 0.05. However, the interaction between black and wages in 1974 (re74) is slightly significant, indicating that, while black is not directly a significant effect on determining employment, there does appear to be a signficant interaction between the black covariate and the wages in 1974 which influences employment. Further research on this interaction could be interesting to further understand social structures.

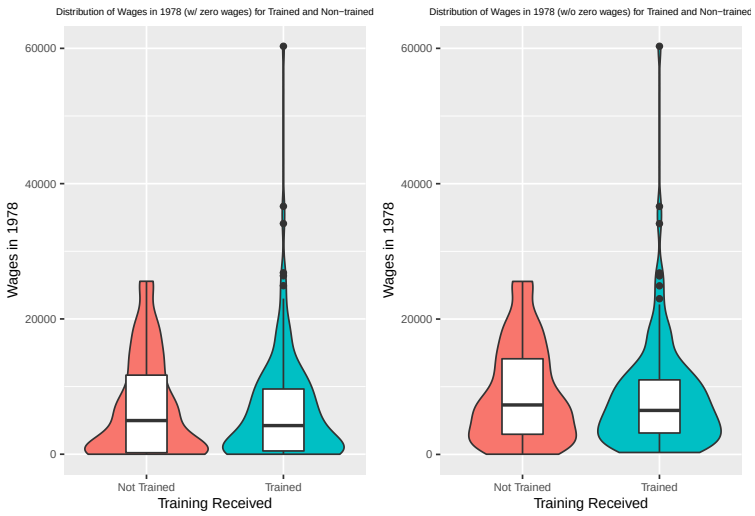
Conclusions

From this model, it appears that treatment does have a positive effect on whether a person had a job in 1978. The confidence interval is quite large, and the accuracy is only about 61.1% accurate at predicting our data. However, this model can still provide insight into whether job training can have an effect on whether a low wage male will be employed in 1978.

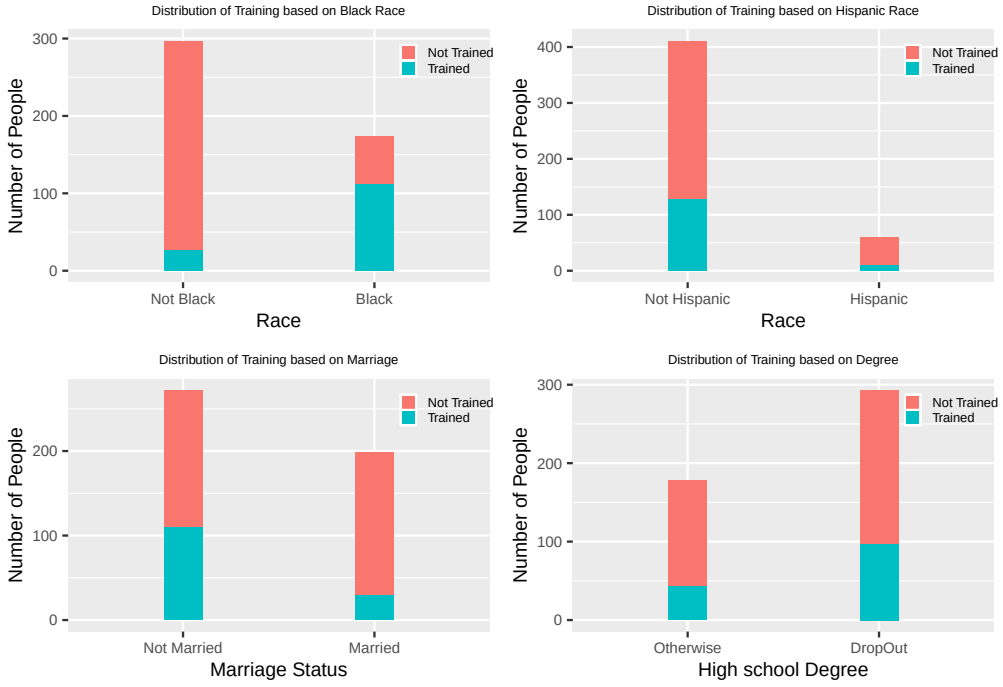
For further research, an interesting association appears to exist between age and treatment; further research to determine why this interaction exists could help to target the treatment more effectively to different age groups.

APPENDIX

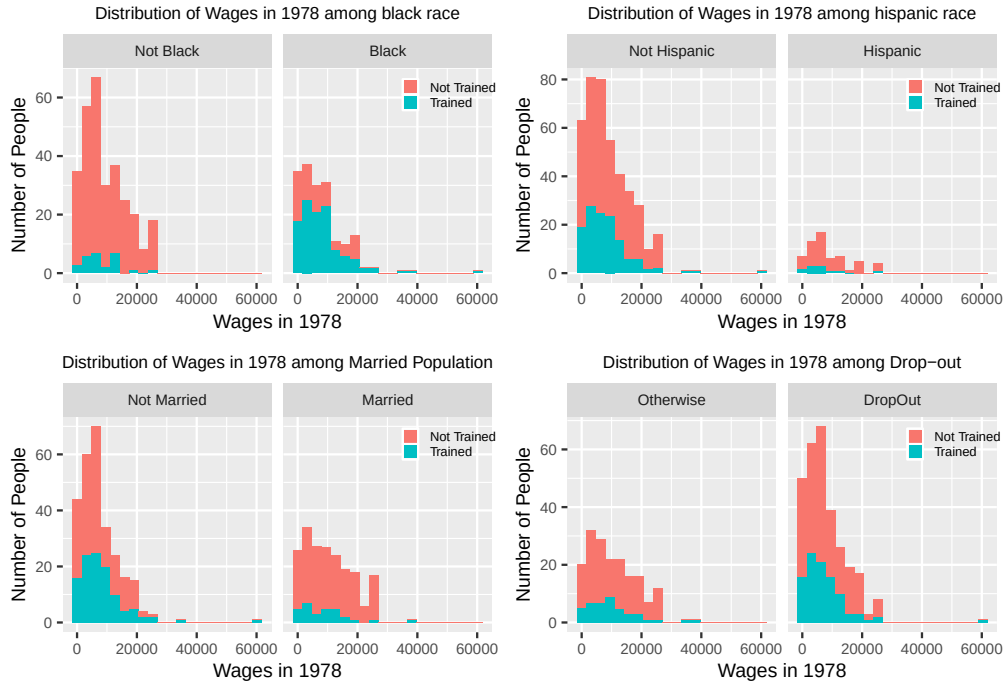
- **Graph 1**



- **Graph 2**



- **Graph 3**




```
## Analysis of Deviance Table
##
## Model 1: emp78 ~ age_c + re74_c + black + treat + re74_c:black + age_c:treat +
##     black:treat
## Model 2: emp78 ~ age_c + re74 + black + treat + re74:black + age_c:treat
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       606     621.27
## 2       607     625.26 -1  -3.9947  0.04564 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table
##
## Model 1: emp78 ~ age_c + re74_c + black + treat + re74_c:black + age_c:treat +
##     black:treat
## Model 2: emp78 ~ age_c + re74 + black + treat + age_c:treat + black:treat
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       606     621.27
## 2       607     625.19 -1    -3.92  0.04771 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table
##
## Model 1: emp78 ~ age_c + re74_c + black + treat + re74_c:black + age_c:treat +
##     black:treat
```

```
## Model 2: emp78 ~ age_c + re74 + treat + black:treat + black:re74 + age_c:treat
##   Resid. Df Resid. Dev Df    Deviance Pr(>Chi)
## 1       606     621.27
## 2       606     621.27  0 -1.1369e-13
```