Effects of Job Training on Wages Part 1

Team 3 Green

Abhijith Tammanagari, Clarissa Ache, Marlyne Hakizimana, Shufan Xia, Tigran Harutyunyan

DATA & PROJECT OBJECTIVE

- Data for this analysis includes:

  1. The **treatment group** (those who received the training) includes male participants within a subset from NSW data for which 1974 earnings can be obtained
  2. The **control group** (those who did not receive the training) includes all the unemployed males in 1976 whose income in 1975 was below the poverty level. Not obtained as part of the NSW experiment.

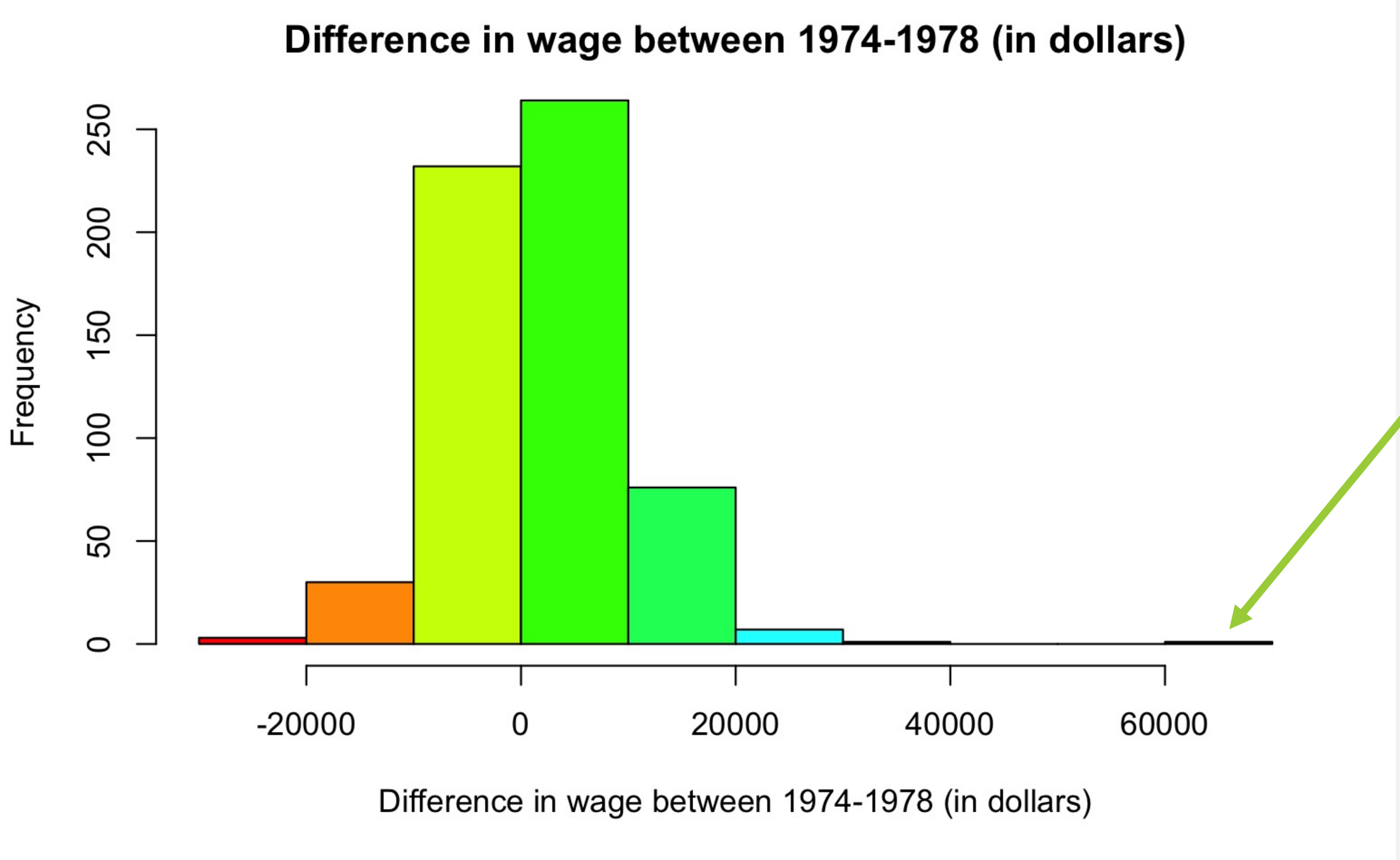
- Variables obtained:

Variable Name	**Type**	**Description**
treat	Binary	1 if participant received job training, 0 if participant did not receive job training.
age	Numeric, discrete	age in years
educ	Numeric, discrete	years of education
black	Binary	1 if race is black, 0 otherwise.
hisp	Binary	1 if Hispanic ethnicity, 0 otherwise.
married	Binary	1 if married, 0 otherwise.
nodegree	Binary	1 if participant dropped out of high school, 0 otherwise.
differ7874	Numeric, continuous (response variable)	real annual earnings difference from 1978 -1974.

- **We will fit a multiple linear regression model to this data to determine if there is evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training.**

ABOUT THE DATA: RESPONSE VARIABLE

- Response variable roughly follows a normal distribution



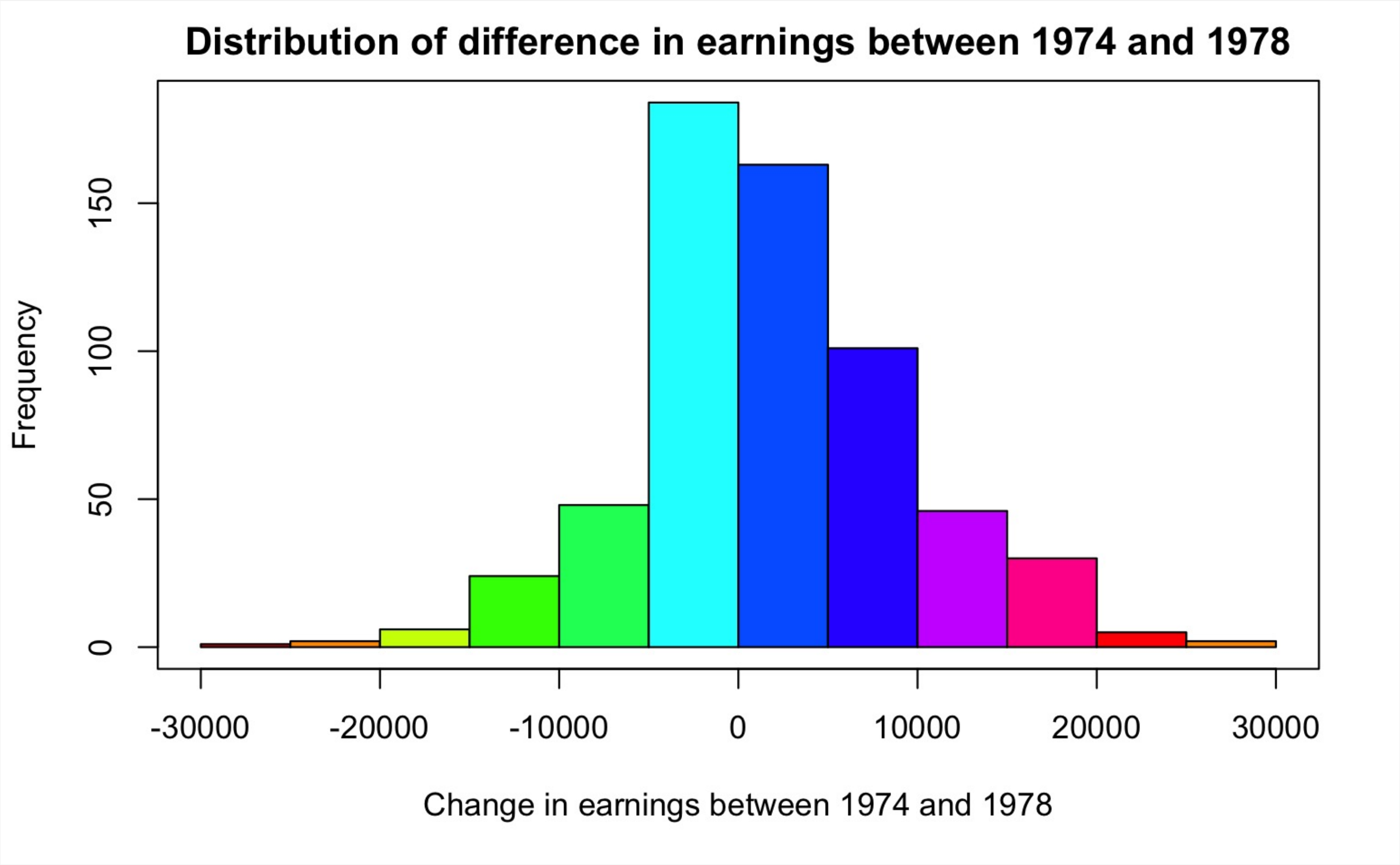
On average, the Difference in wage for all observation is $ 2,235.29
Min = - $ 25,256.80
Max = $ 60,307.93

Only one observations has a Wage difference between 60,000 and 70,000

If we remove this observation, the curve looks more "normal"

ABOUT THE DATA: RESPONSE VARIABLE

- Response variable roughly follows a normal distribution



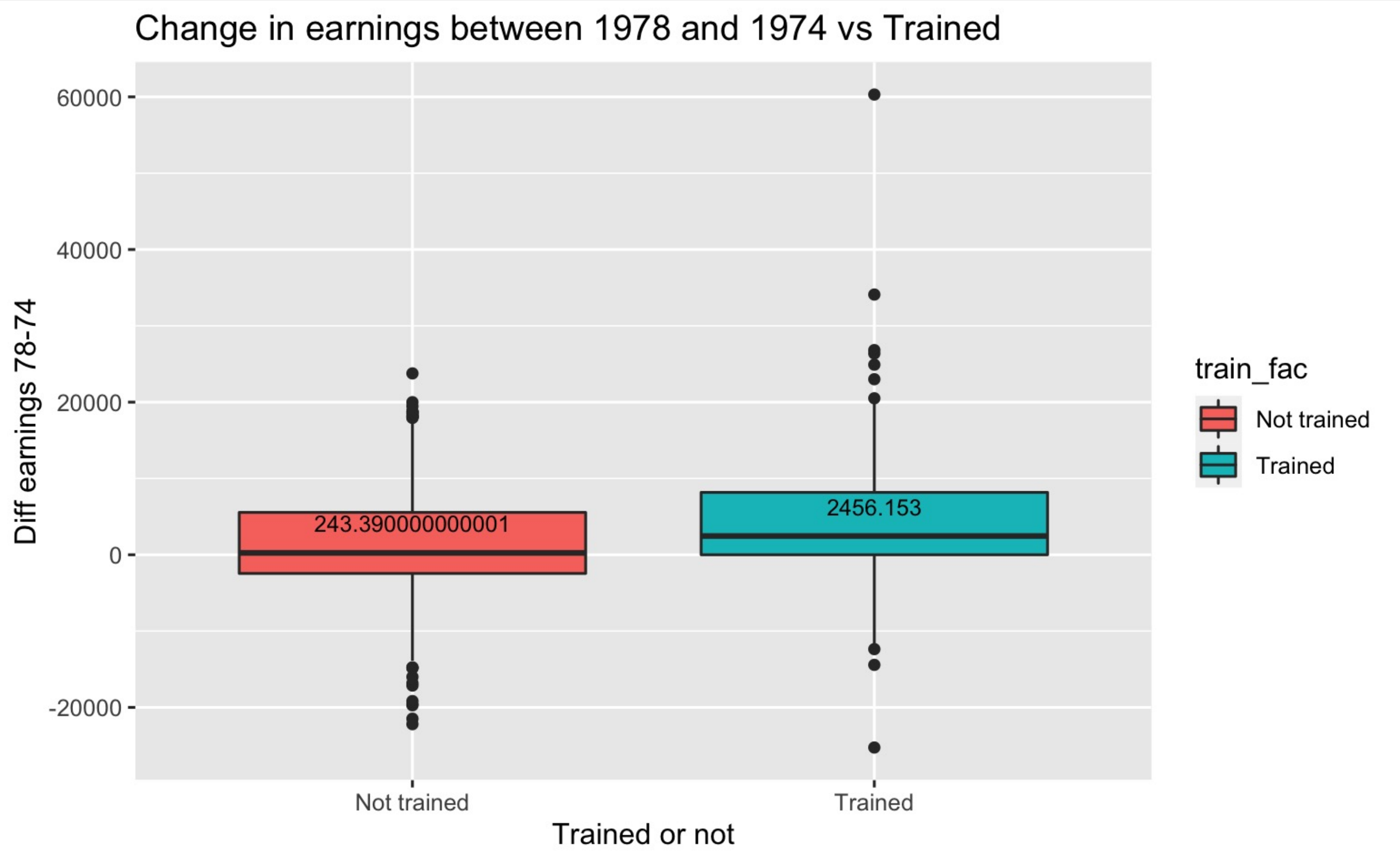
We did a quick test of how the plot would look without the one extreme observation

This looks more "Normal"

However, we fit the model to the full data because we found out (later) that the influence of this point is not high enough to remove as an outlier, and it virtually doesn't change the model.

ABOUT THE DATA: PREDICTORS

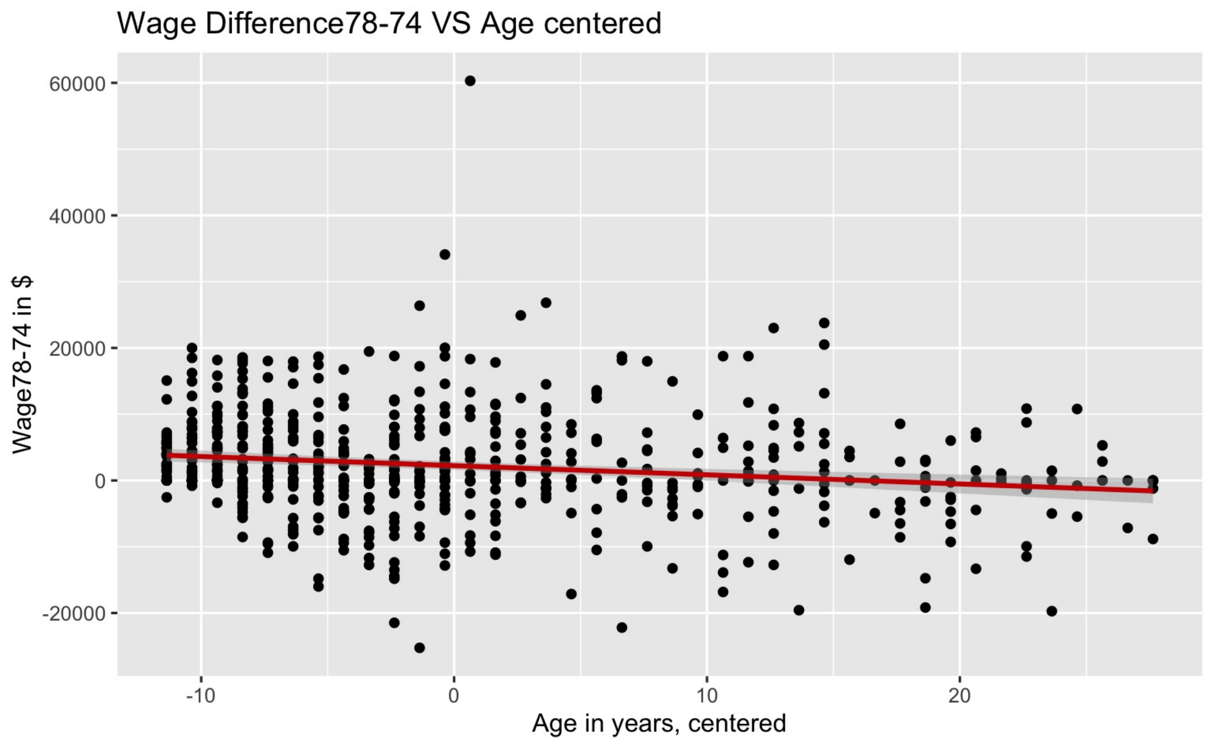
- We expect the treatment (or training) variable to have an effect on the difference of wages

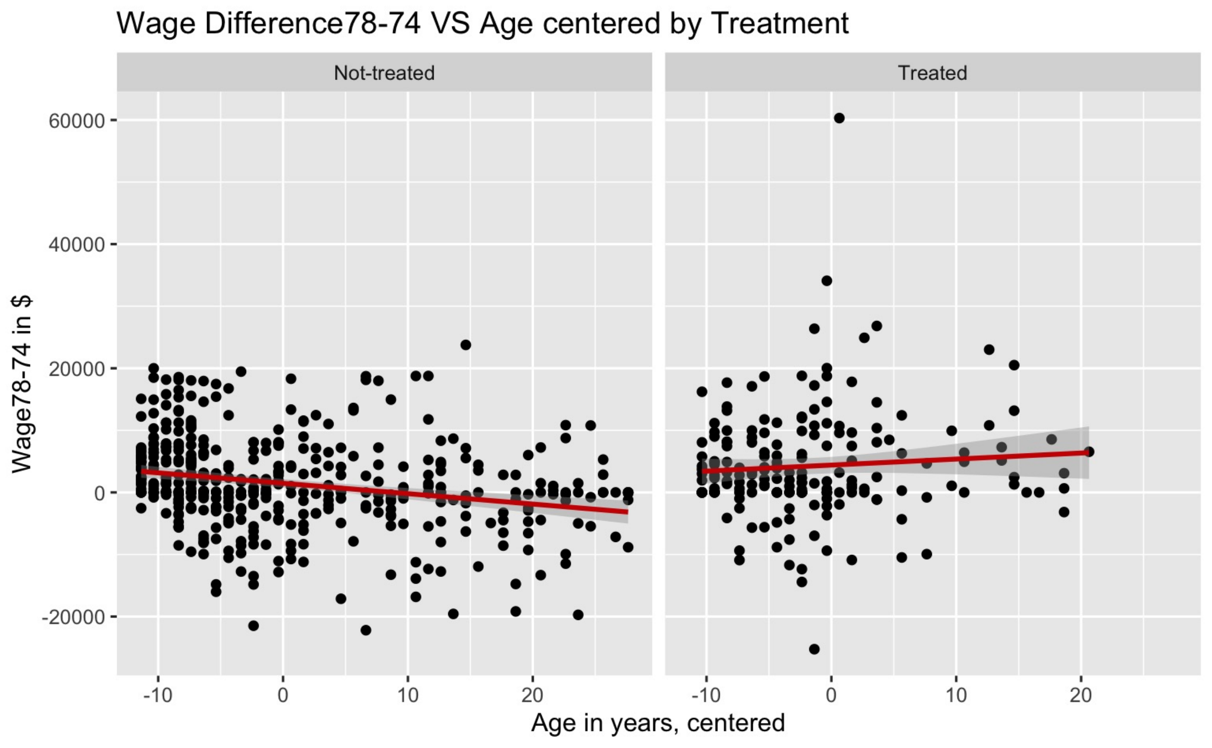


The label on the plot is the median for each level

ABOUT THE DATA: PREDICTORS

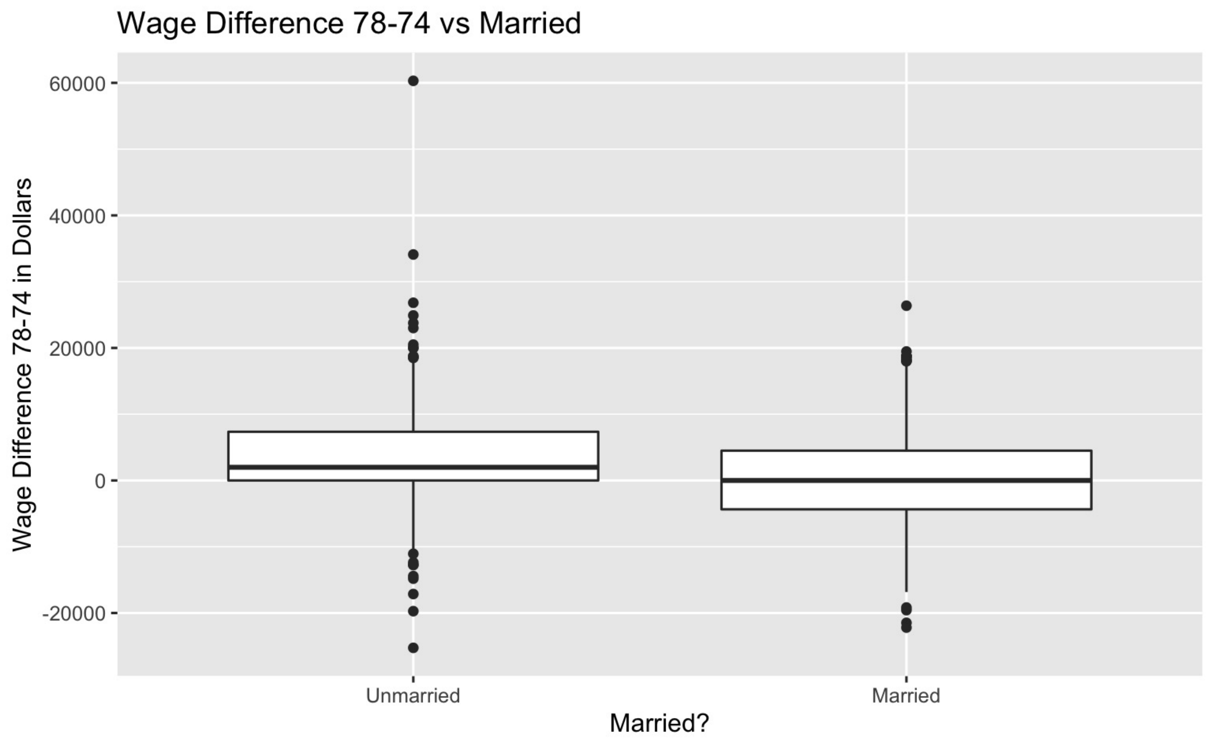
- Wage difference seems to decrease with age in general (right) . However, for trained employees, the trend is positive! (left)

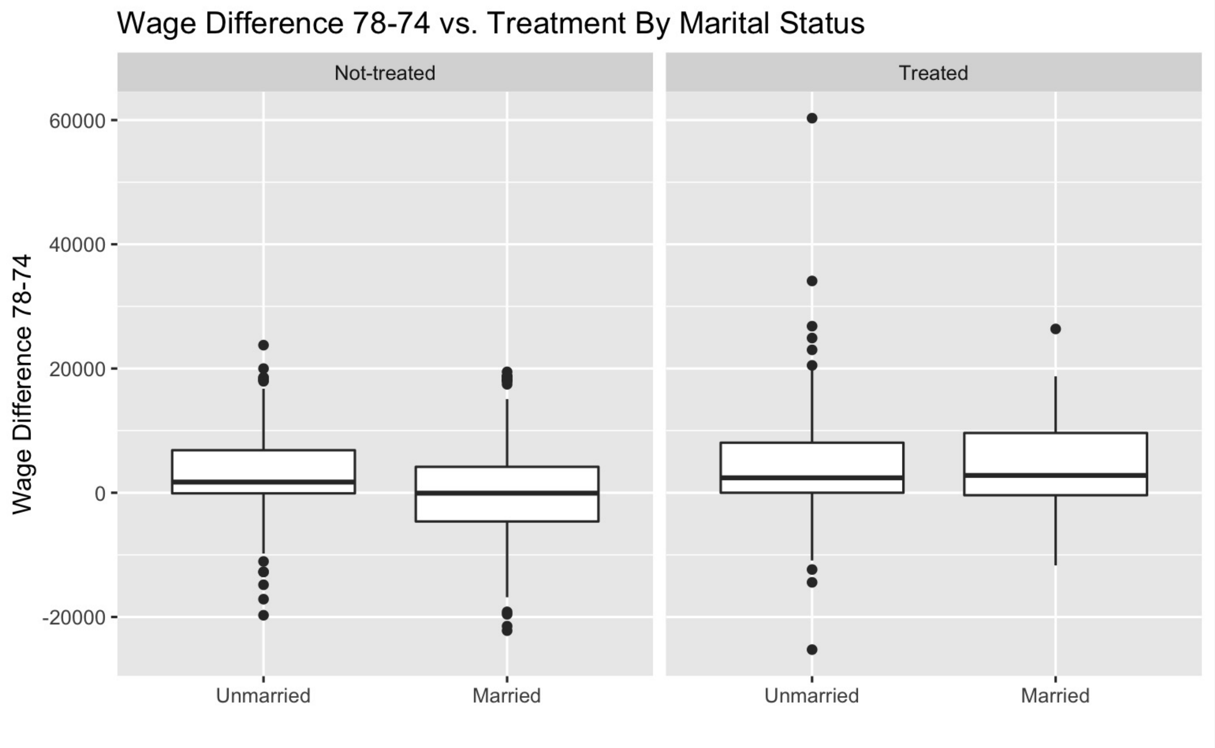




ABOUT THE DATA: PREDICTORS

- Wage difference for Married men seems to be less than it is for unmarried men, in general.
- However, trained married men have slightly higher difference in wage than unmarried trained men.





MODEL BUILDING & VALIDATION

Model Building

- Used stepwise with AIC to simplify the model because we are using this model to understand relationships between variables, not necessarily predict an outcome
- Performed Chi-squared test to test significance of race variables (Black and Hispanic), years of education, non-degree and their interactions with the Treatment variable.
  - None are significant, therefore, they were excluded from the model
- Checked high leverage points. There are a few, however, none have high influence (Cook's distance). No observations were removed as outliers.
- Multicollinearity was also checked. Variables included in the model have low correlation.
- No transformations were required for any of the variables, except for the Age (it was centered). Linearity, normality, equal variance, and independence assumptions are plausible for our final model.

Model Validation:

- R-squared of the final model is %8

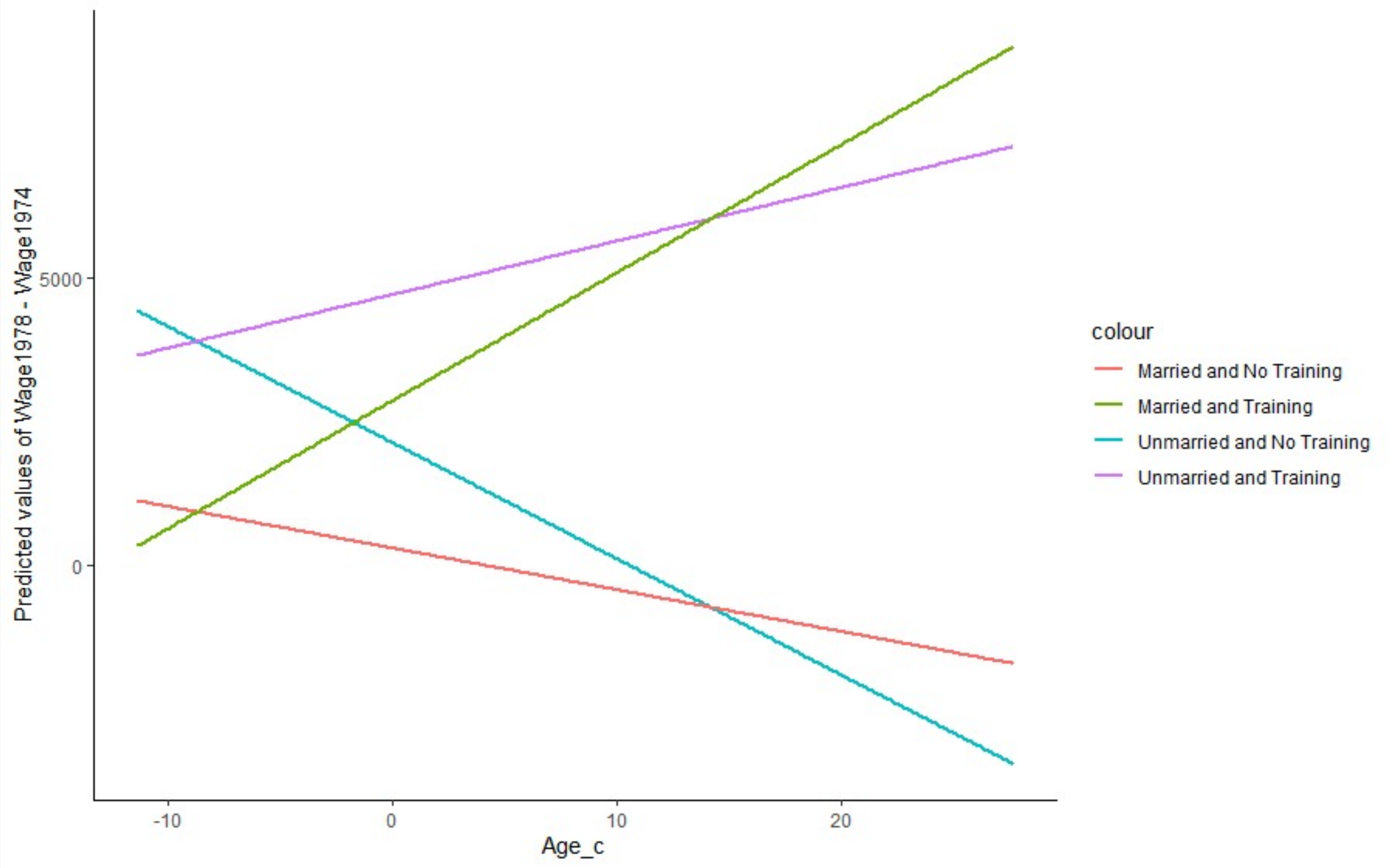
MODEL & FINDINGS

$$Diff7874 = \beta_0 + \beta_1 Treatment + \ \beta_2 Age + \beta_3 Married + \beta_4 Treatment\!:\!Age + \beta_5 Age\!:\!Married + \ \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

Model Output:

Coefficients	Estimate	Std. Error	P-values		2.50%	97.50%
(Intercept)	2122.28	533.38	7.76E-05	***	1074.79	3169.76
Treatment=1	2572.97	727.98	0.00044	***	1143.31	4002.63
Age (centered)	-201.94	51.77	0.000107	***	-303.61	-100.28
Married=1	-1833.72	715.59	0.010631	*	-3239.05	-428.40
Treated=1:Age(centered)	294.99	89.63	0.001056	**	118.96	471.02
Age(centered):Married=1	129.44	70.76	0.067862	.	-9.53	268.41

AGE : MARRIAGE & AGE : TRAINING INTERACTIONS

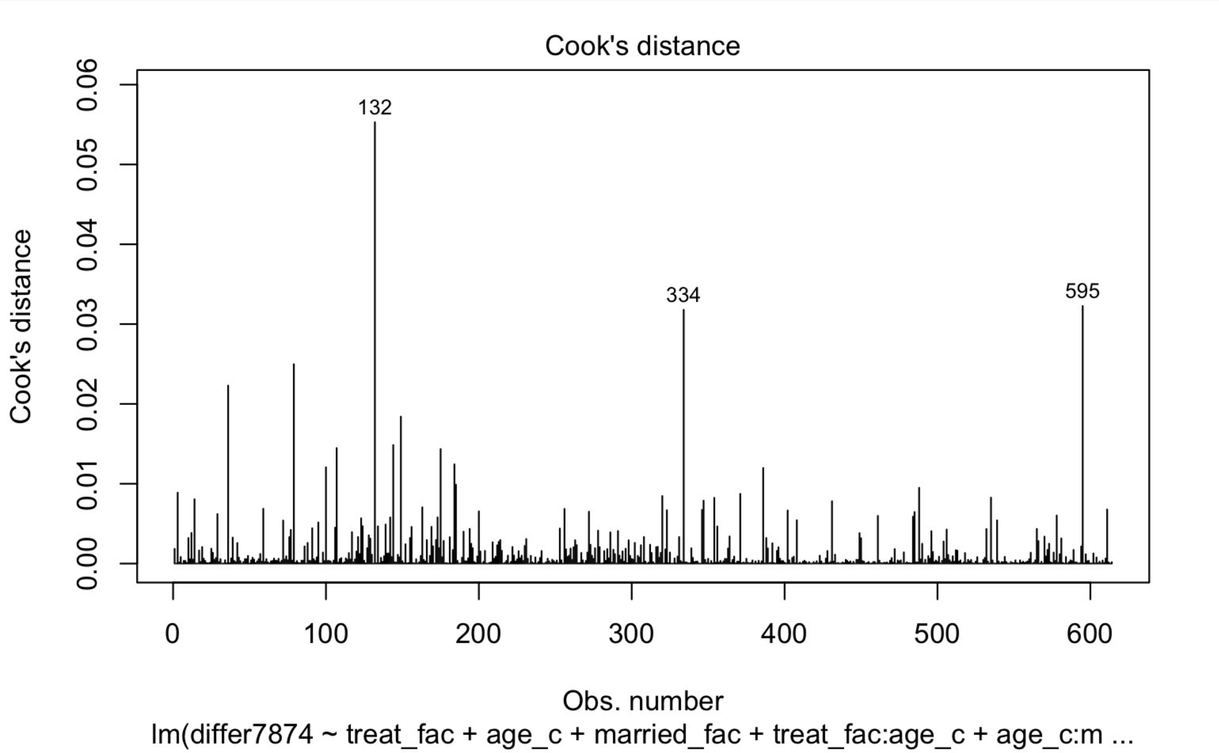


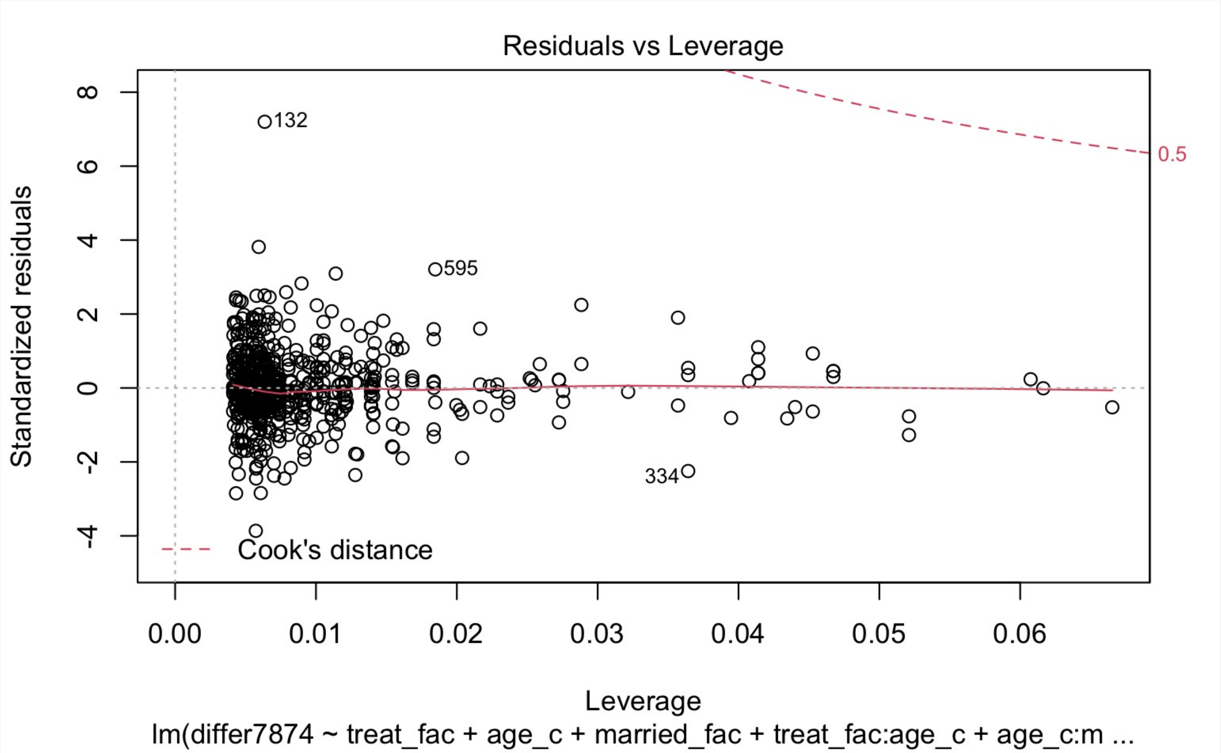
MODEL LIMITATIONS & NEXT STEPS

- Number of Observations:
  - Low counts for Hispanic. Limits the ability to look at interactions.
  - Low counts of Married AND Treated employees (only 35 observations). Limits the model fit.
- When did employees received the training is also likely to affect its effectiveness (and impact in wage difference).
  - If an employee was trained at the beginning of 1975, it is more likely this person came to see the realization of this investment by 1978, compared to a person who got trained by the end of 1977, right before the ”final” income was measured.
- What if people got married during the training period (1975-1978), are they married or non-married? With the average age of observed employees being 27, it is plausible this has an impact in the model fit.
- Looking at the residuals of the final model, although they do look random, they seem to be concentrated in two levels (Y=0, and Y=~4000), which suggests that a binary variable for the Wage difference may be better for explaining the behavior of this data.

OUTLIERS

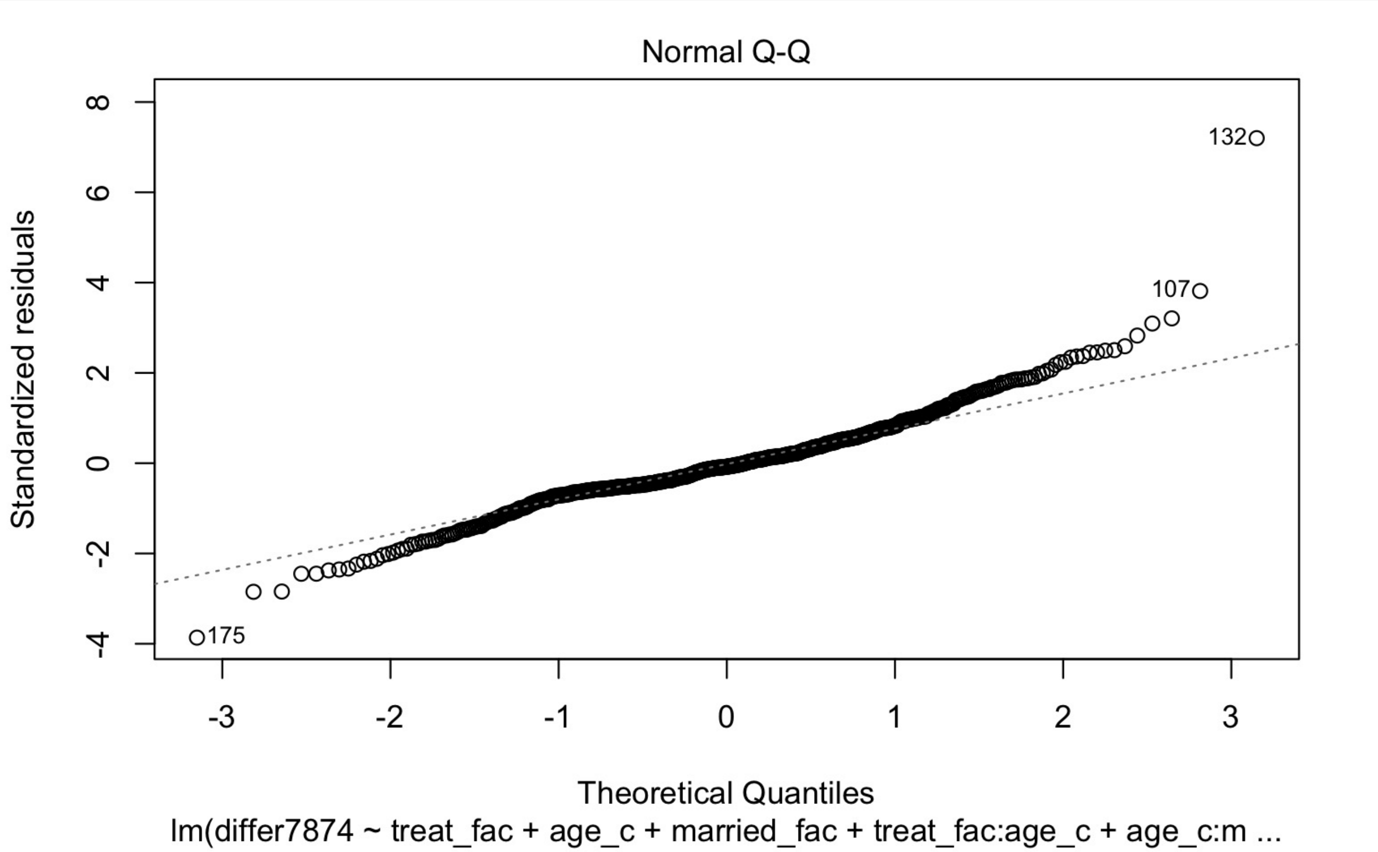
- As expected, the one point with highest influence is the one with the largest salary difference.
- The Cook's Distance for this all points is still below the 0.5 threshold





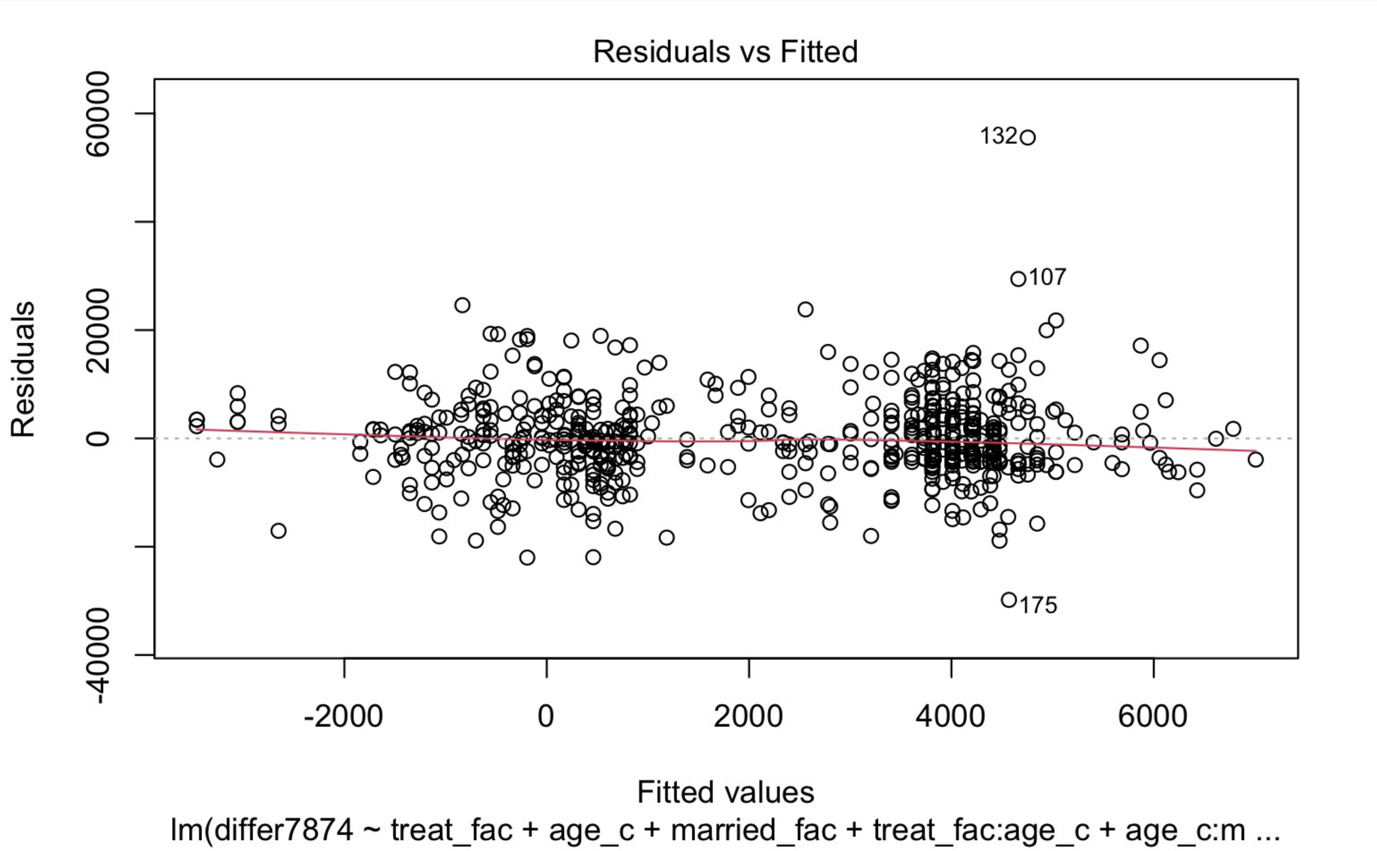
MODEL ASSESSMENT

- Again, the response variable looks roughly Normal. Normality assumption is still plausible.



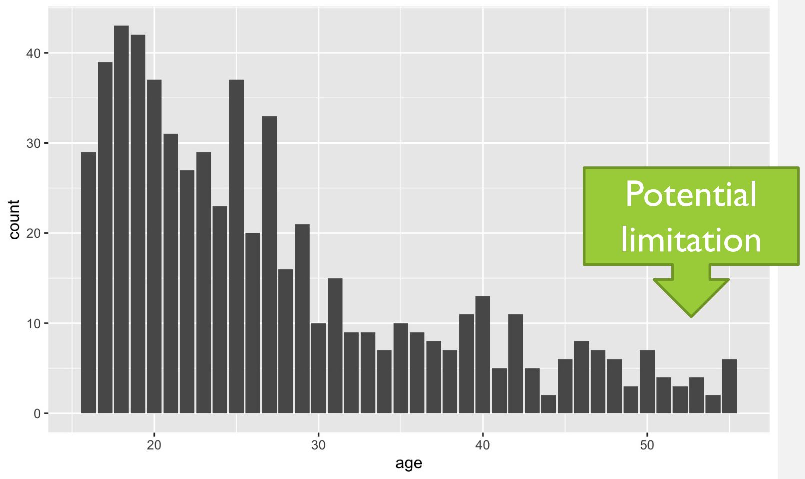
MODEL ASSESSMENT

- Looking at the residuals of the final model, although they do look random, they seem to be concentrated in two levels (Y=0, and Y=~4000), which suggests that a binary variable for the Wage difference may be better for explaining the behavior of this data.

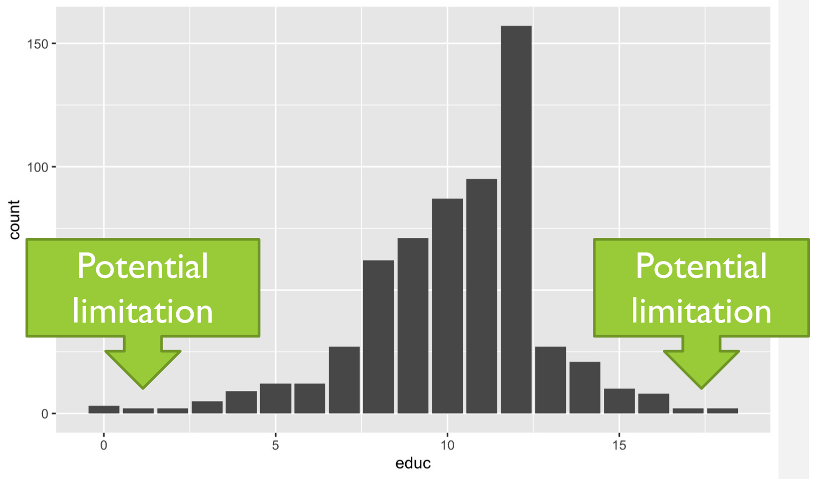


ABOUT THE DATA: OBSERVATION COUNTS

- We checked the number of observations for each variable's levels.

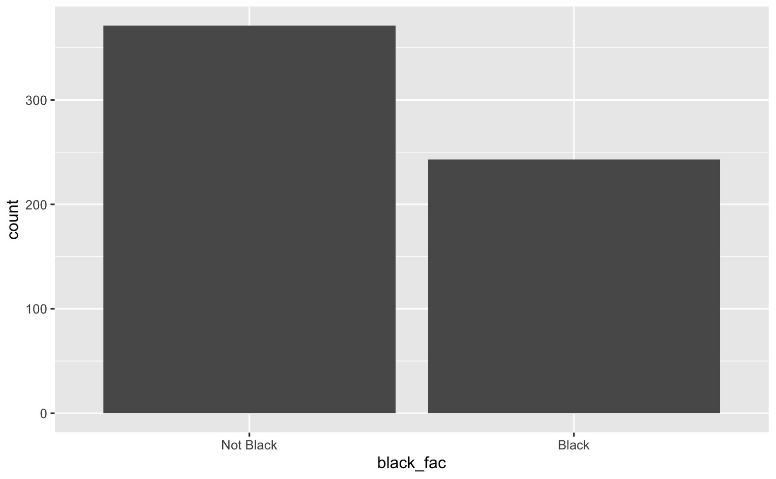


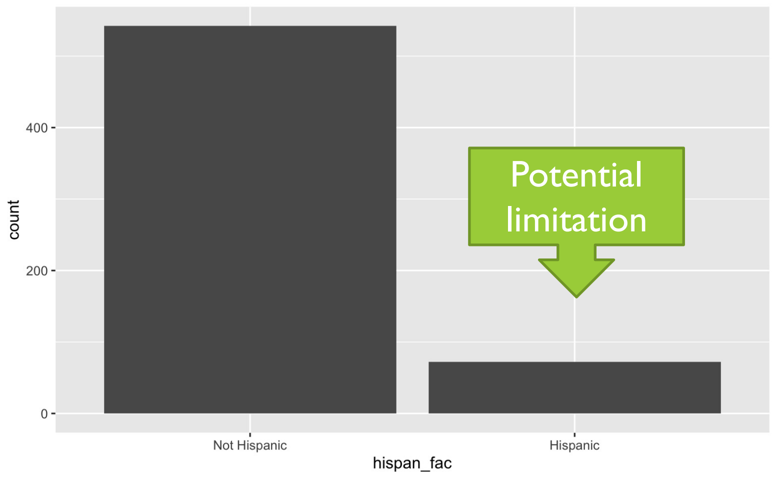
Potential limitation



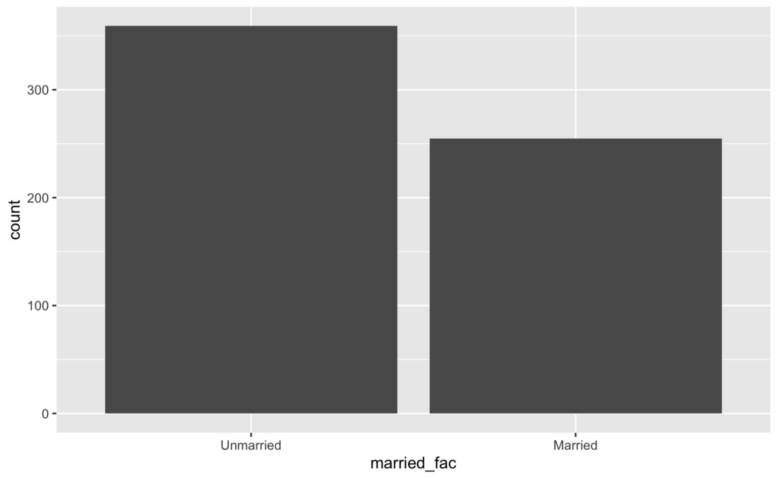
Potential limitation

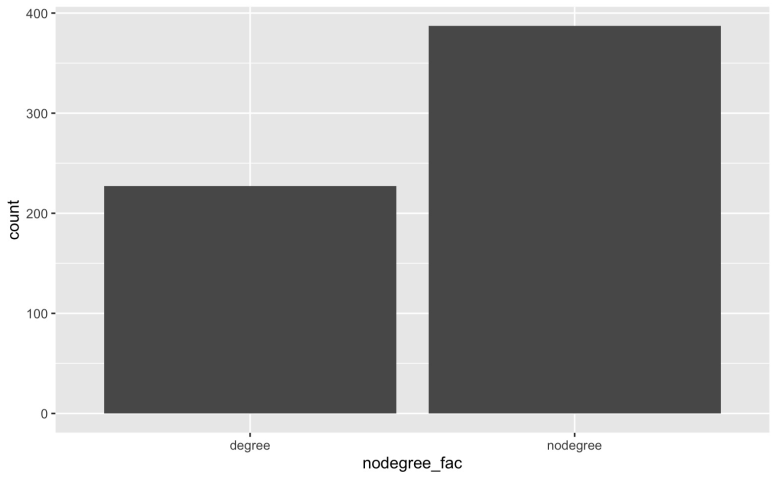
Potential limitation





Potential limitation


count
Unmarried
Married
married_fac


count
degree
nodegree
nodegree_fac