Team Project 1

Allison Young and Anna Berman

10/24/2018

Data Overview

In the 1970s, researchers in the United States ran several randomized experiments intended to evaluate public policy programs. One of the most famous experiments is the National Supported Work Demonstration (NSWD), in which researchers wanted to assess whether or not job training for disadvantaged workers had an effect on their wages. Eligible workers were randomly assigned either to receive job training or not to receive job training. Since this is a randomized experiment, we can make causal claims about the effect of job training on wages for this population of workers.

We analyze a subset of the data from the NSWD. These and other data were originally analyzed in a highly influential paper by the economist Robert Lalonde. The reference for the study is Lalonde, R. J. (1986), Evaluating the econometric evaluations of training programs with experimental data, The American Economic Review, 76, 604 - 620.

We will use linear and logistic regression modeling to answer the following questions of interest.

- Is there evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training? What is a likely range for the effect of training? Is there any evidence that the effects differ by demographic groups? Are there other interesting associations with wages that are worth mentioning?
- 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? What is a likely range for the effect of training? Is there any evidence that the effects differ by demographic groups? Are there other interesting associations with positive wages that are worth mentioning?

A summary of the dataset used in both our linear and logisitic regessions is summarized below:

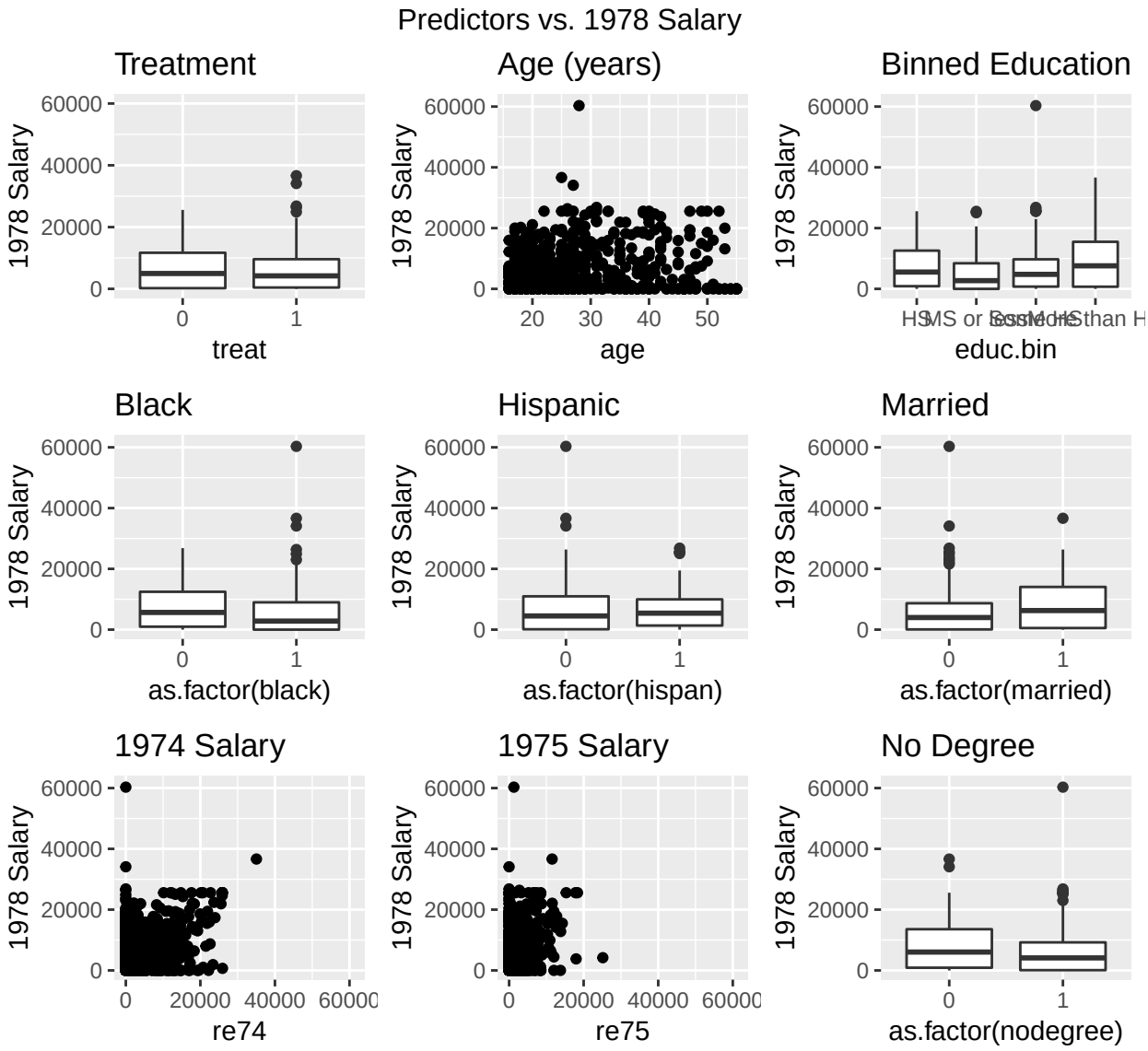
```
##        X       treat        age             educ           black
##  NSW1   :  1   0:429   Min.   :16.00   Min.   : 0.00   Min.   :0.0000
##  NSW10  :  1   1:185   1st Qu.:20.00   1st Qu.: 9.00   1st Qu.:0.0000
##  NSW100 :  1           Median :25.00   Median :11.00   Median :0.0000
##  NSW101 :  1           Mean   :27.36   Mean   :10.27   Mean   :0.3958
##  NSW102 :  1           3rd Qu.:32.00   3rd Qu.:12.00   3rd Qu.:1.0000
##  NSW103 :  1           Max.   :55.00   Max.   :18.00   Max.   :1.0000
##  (Other):608
##      hispan          married          nodegree           re74
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :    0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:    0
##  Median :0.0000   Median :0.0000   Median :1.0000   Median : 1042
##  Mean   :0.1173   Mean   :0.4153   Mean   :0.6303   Mean   : 4558
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 7888
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :35040
##
##       re75             re78            re78c           re75c
##  Min.   :    0.0   Min.   :    0.0   Min.   :-6793   Min.   :-2185
##  1st Qu.:    0.0   1st Qu.:  238.3   1st Qu.:-6555   1st Qu.:-2185
##  Median :  601.5   Median : 4759.0   Median :-2034   Median :-1583
##  Mean   : 2184.9   Mean   : 6792.8   Mean   :    0   Mean   :    0
##  3rd Qu.: 3249.0   3rd Qu.:10893.6   3rd Qu.: 4101   3rd Qu.: 1064
```

```
##  Max.   :25142.2   Max.   :60307.9   Max.   :53515   Max.   :22957
##
##      re74c            agec                 educ.bin
##  Min.   :-4558   Min.   :-11.363   HS          :157
##  1st Qu.:-4558   1st Qu.: -7.363   MS or less  :134
##  Median :-3515   Median : -2.363   Some HS     :253
##  Mean   :    0   Mean   :  0.000   More than HS: 70
##  3rd Qu.: 3331   3rd Qu.:  4.637
##  Max.   :30483   Max.   : 27.637
##
```

Linear Regression

Exploratory Data Analysis

For concern of multicollinearity, we cannot include both nodegree and education in our model (nodegree is in essence a binned version of education with 0 being over 12 years of education and 1 being less than 12 years of education). We were originally concerned with including both 1974 salary (re74) and 1975 salary (re75), however, the correlation between these two variables is only 0.55 which low enough to allow both in our model. No other variables had high enough correlation to be a multicollinearity concern.



Model Selection

We evaluated each model based on model's R-squared and whether addition variables and interactions resulted in a significant or near-significant nested F test. Through a series of modeling fittings, we examined a variety of linear models to answer the question, 'Is there evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training?'. We attempted logging our outcome variable, 1978 salary (re78), using nodegree as opposed to education, and using education as a continuous variable as well as a binned factor variable. We also looked at potential interaction effects between treatment and education as well as treatment and black (see Appendix Fig. 1 for plots of potential interaction effects). Additionally, we used mean-centered continuous variables to aid in interpretation.

Ultimately, we selected the following model.

The residuals of this model are normally distributed and fit our assumptions of linear regression (see Appendix Fig. 2 for residual plots).

Influential points (see Appendix Fig. 3)

```
##
## Call:
## lm(formula = re78 ~ treat + agec + educ.bin + black + hispan +
##     married + re74c + re75c, data = lalonde)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13858  -4842  -1516   4062  54869
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6655.8106   724.1249   9.192  < 2e-16 ***
## treat1                1612.9207   779.9046   2.068   0.0391 *
## agec                     6.9456    32.4362   0.214   0.8305
## educ.binMS or less   -1693.6195   844.2911  -2.006   0.0453 *
## educ.binSome HS         29.6427   726.9096   0.041   0.9675
## educ.binMore than HS  2254.6884  1003.6187   2.247   0.0250 *
## black                -1278.3191   767.4450  -1.666   0.0963 .
## hispan                 357.4441   931.7120   0.384   0.7014
## married                518.6474   696.6006   0.745   0.4568
## re74c                    0.3044     0.0580   5.247 2.14e-07 ***
## re75c                    0.2205     0.1043   2.113   0.0350 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6934 on 603 degrees of freedom
## Multiple R-squared:  0.1526, Adjusted R-squared:  0.1385
## F-statistic: 10.86 on 10 and 603 DF,  p-value: < 2.2e-16
```

```
##                           Estimate    Std.Error          2.5%        97.5%
## (Intercept)           6655.8106348 7.241249e+02  5.233697e+03 8077.9238346
## treat1                1612.9206555 7.799046e+02  8.126151e+01 3144.5797973
## agec                     6.9455831 3.243616e+01 -5.675599e+01   70.6471528
## educ.binMS or less   -1693.6194714 8.442911e+02 -3.351728e+03  -35.5112419
## educ.binSome HS         29.6427548 7.269096e+02 -1.397939e+03 1457.2247332
## educ.binMore than HS  2254.6883702 1.003619e+03  2.836756e+02 4225.7010937
## black                -1278.3190756 7.674450e+02 -2.785509e+03  228.8707398
## hispan                 357.4440745 9.317120e+02 -1.472351e+03 2187.2387178
## married                518.6474110 6.966006e+02 -8.494106e+02 1886.7054177
## re74c                    0.3043451 5.800401e-02  1.904307e-01    0.4182595
## re75c                    0.2205303 1.043544e-01  1.558821e-02    0.4254725
```

Interpretation

Intercept: For non-black, non-hispanic, un-married individuals of average age, average 1974 and 1975 salaries, with High School only education, who did not recieve treatment, we estimate the average salary in 1978 to be $6655.81 (95% CI: $5233.7, $8077.92).

Treatment: Holding all else constant, individuals who participated in the treatment are estimated to have average 1978 salaries increased by $1612.92 (95% CI: $81.26, $3144.58).

Age: Holding all else constant, for each 10 years an individual ages on average we estimate his salary to increase by $69.46 (95% CI: $-567.56, $706.47). Given that this confidence interval includes 0, we are not confident that there is a meaningful effect of age on 1978 salary.

Education: Holding all else contant, for an individual with:

- Less than a middle school education: we estimate avergage 1978 salary to be $1693.62 less (95% CI: $-3351.73, $-35.51).
- Some high school education: we estimate avergage 1978 salary to be $29.64 more (95% CI: $-1397.94, $1457.22). Given that this confidence interval includes 0, we are not confident that there is a meaningful effect of some high school compared to completion of high school on 1978 salary.
- More than a high school education: we estimate avergage 1978 salary to be $2254.69 more (95% CI: $283.68, $4225.7).

Black: Holding all else constant, for black individuals we estimate average 1978 salaries to be $1278.32 less (95% CI: $-2785.51, $228.87). Given that this confidence interval includes 0, we are not confident that there is a meaningful effect of blackness on 1978 salary.

Hispanic: Holding all else constant, for black individuals we estimate average 1978 salaries to be $357.44 more (95% CI: $-1472.35, $2187.24). Given that this confidence interval includes 0, we are not confident that there is a meaningful effect of hispanic ethnicity on 1978 salary.

1974 Salary: Holding all else constant, for each $1,000 an individual made in 1974, on average we estimate his 1978 salary to be $304.35 higher (95% CI: $190.43, $418.26).

1975 Salary: Holding all else constant, for each $1,000 an individual made in 1975, on average we estimate his 1978 salary to be $220.53 higher (95% CI: $15.59, $425.47).

Conclusion

Limitations

Appendix

Fig 1: Linear Interaction Plots Fig 2: Linear Residual Plots Fig 3: Influential points