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
```

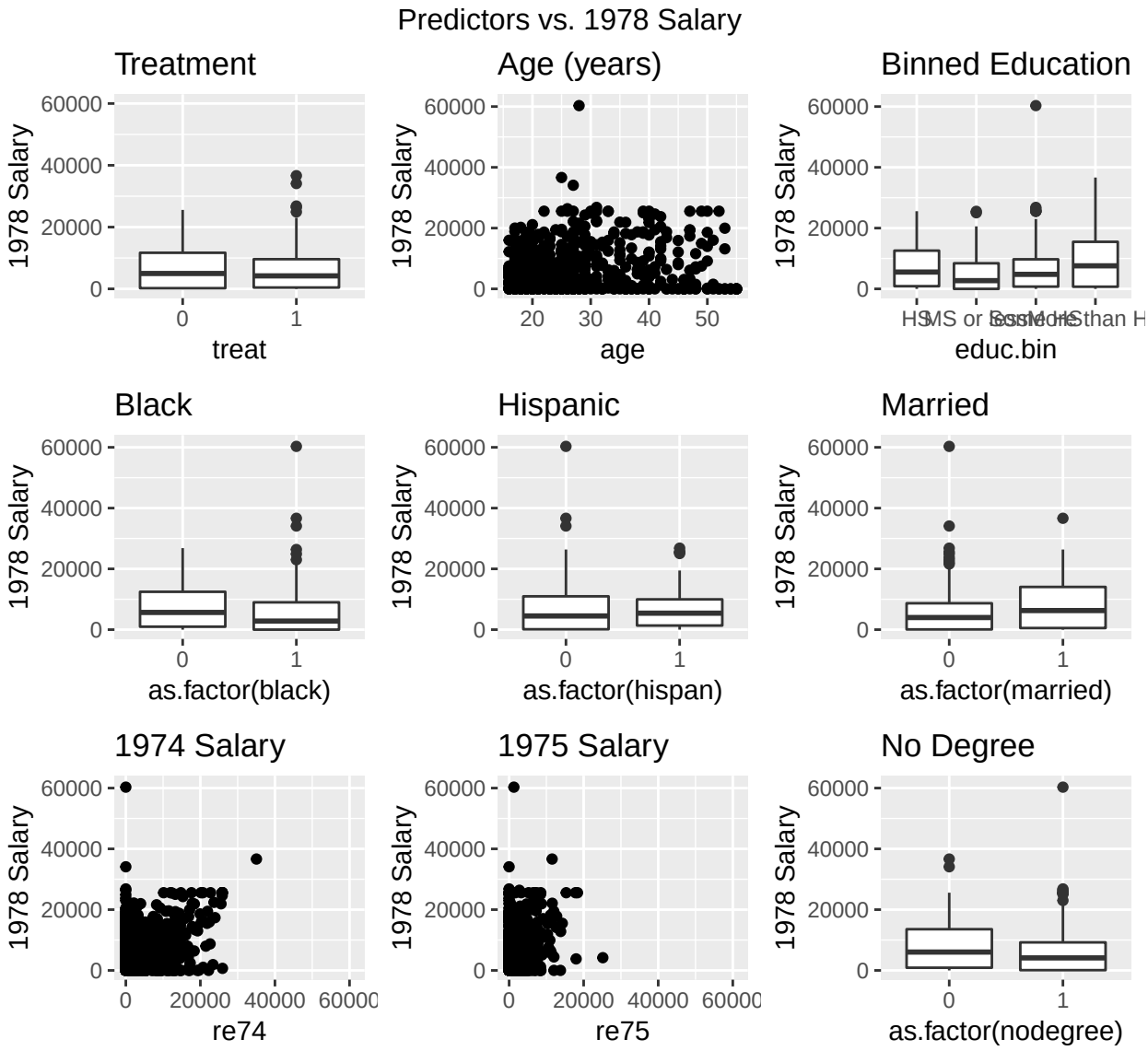
```
##  Max.   :25142.2   Max.   :60307.9   Max.   :53515   Max.   :22957
##
##      re74c             agec           employed78       employed74
##  Min.   :-4558   Min.   :-11.363   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:-4558   1st Qu.: -7.363   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :-3515   Median : -2.363   Median :1.0000   Median :1.0000
##  Mean   :    0   Mean   :  0.000   Mean   :0.7671   Mean   :0.6042
##  3rd Qu.: 3331   3rd Qu.:  4.637   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :30483   Max.   : 27.637   Max.   :1.0000   Max.   :1.0000
##
##          educ.bin        educ.bin2        age2               age3
##  HS          :157   Some HS + :480   Min.   :  0.1319   Min.   :-1467.24
##  MS or less  :134   MS or less:134   1st Qu.: 11.3111   1st Qu.: -399.21
##  Some HS     :253                    Median : 54.2166   Median :  -13.20
##  More than HS: 70                    Mean   : 97.4788   Mean   :  986.08
##                                      3rd Qu.:107.3958   3rd Qu.:   99.69
##                                      Max.   :763.7931   Max.   :21108.80
##
```

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 salary variables as predictors in our model. No other variables had high enough correlation to be a multicollinearity concern.

A plot of each predictor in relation to our outcome variable, 1978 salary is below.



Model Selection

Through a series of modeling fittings, we examined a variety of logistic 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 evaluated each model based on R-squared value and whether addition variables and interactions resulted in a significant or near-significant nested F test results.

We attempted logging our outcome variable (1978 salary (re78)), logging 1974 and 1975 salaries, using nodegree as opposed to education, using education as a continuous variable as well as a binned factor variable. We also looked at potential interaction effects between treatment and education, treatment and black, treatment and hispanic, and treatment and age (see appendix Fig. 1 for plots of potential interaction effects). Additionally, we used mean-centered continuous variables to aid in interpretation.

Before we finalized our model selection we examined the residuals and influential points. The residuals of this model are normally distributed and have constant variance therefore fitting our assumptions of linear regression (see appendix). The most influential points in our model were determined to be corner cases and did not call for alteration of our final model. (For further details on our model's redisuals and influential

points, see appendix).

Ultimately, we selected the model summarized below.

```
##
## 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
```

Interpretation

Our model has an R-squared of 0.15. In other words, our model explains 15% of the variance 1978 salary.

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).

Married: Holding all else constant, for married individuals we estimate average 1978 salaries to be $518.65 more (95% CI: $-849.41, $1886.71). Given that this confidence interval includes 0, we are not confident that there is a meaningful effect of blackness on 1978 salary.

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).

Discussion

Because this is a randomized control trial, we can say treatment may result in increased salaries. However the effect size of treatment may be small.

- The effect is confounded by age, ethnicity, education, marital status, and previous salary.
- Both being black being hispanic, and being married may not have an effect on salary
- 1974 is more representative of earning potential in 1978 compared to 1975
- Unclear whether there is a huge difference between some hs and hs completion. Matters when you have less than ms or more than hs

Limitations

Our model has an R-squared of 0.15. In other words, our model explains 15% of the variance 1978 salary. It seems that we are missing variables in our model that would explain additional variation in 1978 salary, therefore more research is needed to fully understand the relationship between job training programs and salary and the mediating variables in this relationship.

Doesn't work for wealthy or older individuals (Why older?) Also black Outliers

Logistic Regression

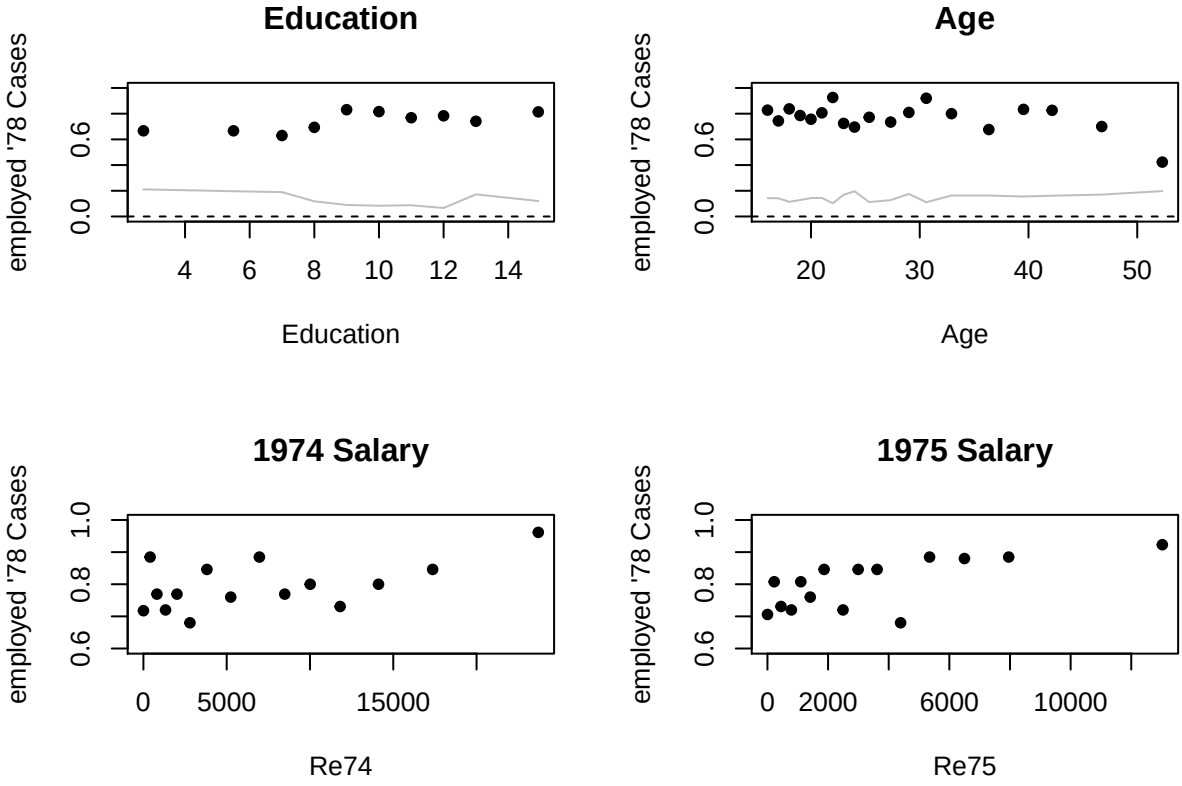
Exploratory Data Analysis

In terms of mulicollinearity, the same reasoning from our linear regression applies to our logistic regression. Therefore our only restriciton is a choice between either education or nodegree.

A plot of each predictor in relation to our outcome variable, employment in 1978 is below (employment being defined as salary above 0).

Table 1: Average employed '78 Cases by predictor

	0	1
treat	0.77	0.76
educ.bin2	0.79	0.67
black	0.80	0.72
hispan	0.76	0.83
married	0.76	0.78
nodegree	0.78	0.76



Model Selection

We examined a variety of linear models to answer the question, '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?'. We evaluated each model based on the area under the curve (AUC) and whether addition variables and interactions resulted in a significant or near-significant change in deviance tests.

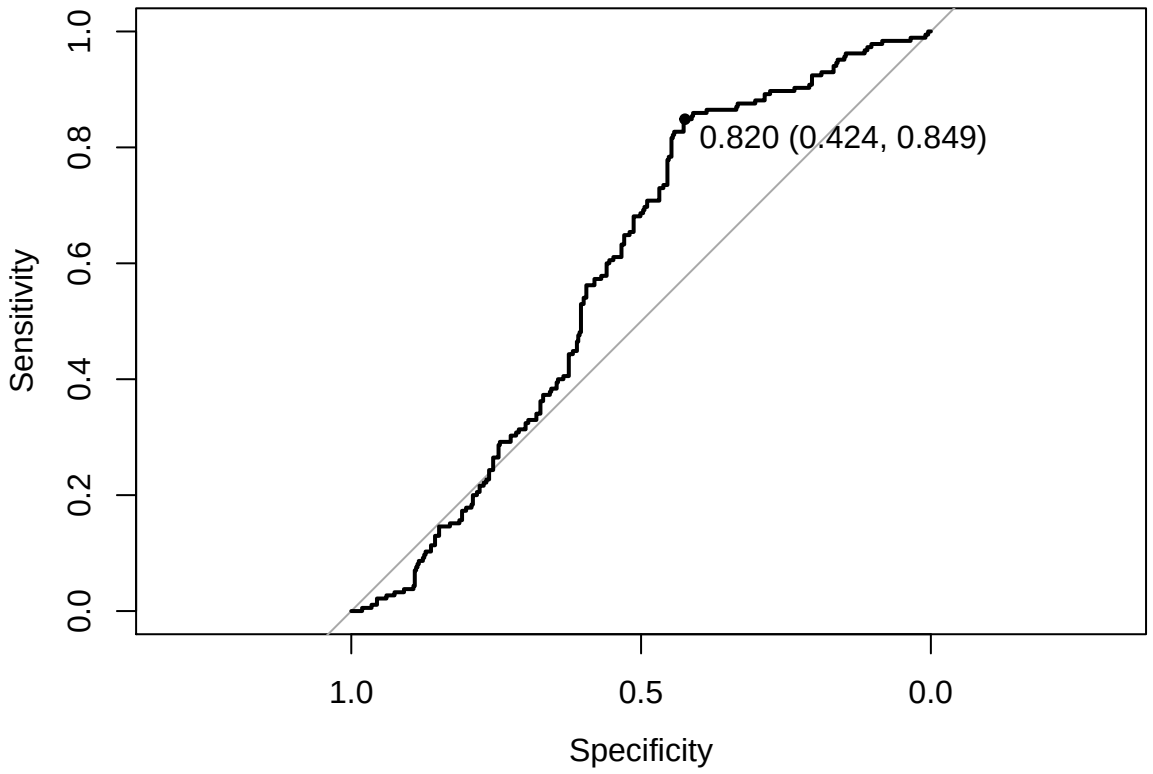
We attempted binning education in multiple ways, nonzero 1974 and 1975 variables. We also examoined potential interactions between treatment and previous salaries as well as interactions between treatment and level of education (see appendix for examination of interaction effects). Additionally, we used mean-centered continuous variables to aid in interpretation.

Before we finalized our model selection we examined the residuals and influential points. The residuals of this model are normally distributed and fit our assumptions of logistic regression (see appendix).

The most influential points in our model were determined to be corner cases and did not call for alteration of our final model. (For further details on our model's redisuals and influential points, see appendix).

Ultimately, we selected the model summarized below.

```
##
## Call:
## glm(formula = employed78 ~ treat * employed74 + agec + married +
##     black + hispan + educ.bin2 + re75c, family = binomial, data = lalonde)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4043   0.3469   0.6049   0.7440   1.4117
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.177e+00  2.796e-01   4.209 2.57e-05 ***
## treat1               7.732e-01  3.244e-01   2.383  0.01716 *
## employed74           4.703e-01  2.771e-01   1.698  0.08959 .
## agec                -3.260e-02  1.039e-02  -3.137  0.00171 **
## married              4.758e-02  2.429e-01   0.196  0.84468
## black               -5.287e-01  2.675e-01  -1.976  0.04810 *
## hispan               2.088e-01  3.601e-01   0.580  0.56201
## educ.bin2MS or less -5.688e-01  2.328e-01  -2.443  0.01456 *
## re75c                1.297e-04  4.511e-05   2.874  0.00405 **
## treat1:employed74   -1.187e+00  4.711e-01  -2.519  0.01176 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 666.5  on 613  degrees of freedom
## Residual deviance: 622.7  on 604  degrees of freedom
## AIC: 642.7
##
## Number of Fisher Scoring iterations: 4
```




```
## [1] "Confusion Matrix"
```

```
##
##     FALSE TRUE
##   0   118   25
##   1   287  184
```

```
## Waiting for profiling to be done...
```

```
## [1] "Confidence intervals"
```

```
##                         Estimate      2.5%     97.5%
## (Intercept)         3.2440128 1.8954738 5.6865021
## treat1              2.1666346 1.1510708 4.1177880
## employed74          1.6005330 0.9269135 2.7524916
## agec                0.9679244 0.9483436 0.9878683
## married             1.0487287 0.6527574 1.6943581
## black               0.5893666 0.3481850 0.9953958
## hispan              1.2321777 0.6250256 2.5907711
## educ.bin2MS or less 0.5662319 0.3598939 0.8980752
## re75c               1.0001297 1.0000460 1.0002234
## treat1:employed74   0.3051571 0.1219336 0.7774356
```

Interpretation

Our model has an AUC of 0.59. Using the suggested threshold of 0.82, our model has a sensitivity of 0.391 and a specificity of 0.175. In other words, our model correctly predicts 39.1% of nonzero wage earners and 82.5% of zero wage earners.

Intercept: For non-black, non-hispanic, un-married individuals of average age, average 1975 salaries and a zero 1974 salary, with some High School or more education, who did not recieve treatment, we estimate the

odds of nonzero salary 1978 to be $3.24 (95% CI: $1.895, $5.69).

Treatment: Holding all else constant, for individuals who participated in the treatment we estimate the odds of nonzero wage in 1978 to increase by a factor of 2.17 (95% CI: 1.15, 4.12).

Education: Holding all else contant, for an individual less than a middle school education we estimate e estimate the odds of nonzero wage in 1978 to decrease by a factor of 0.57 less (95% CI: 0.36, 0.9).

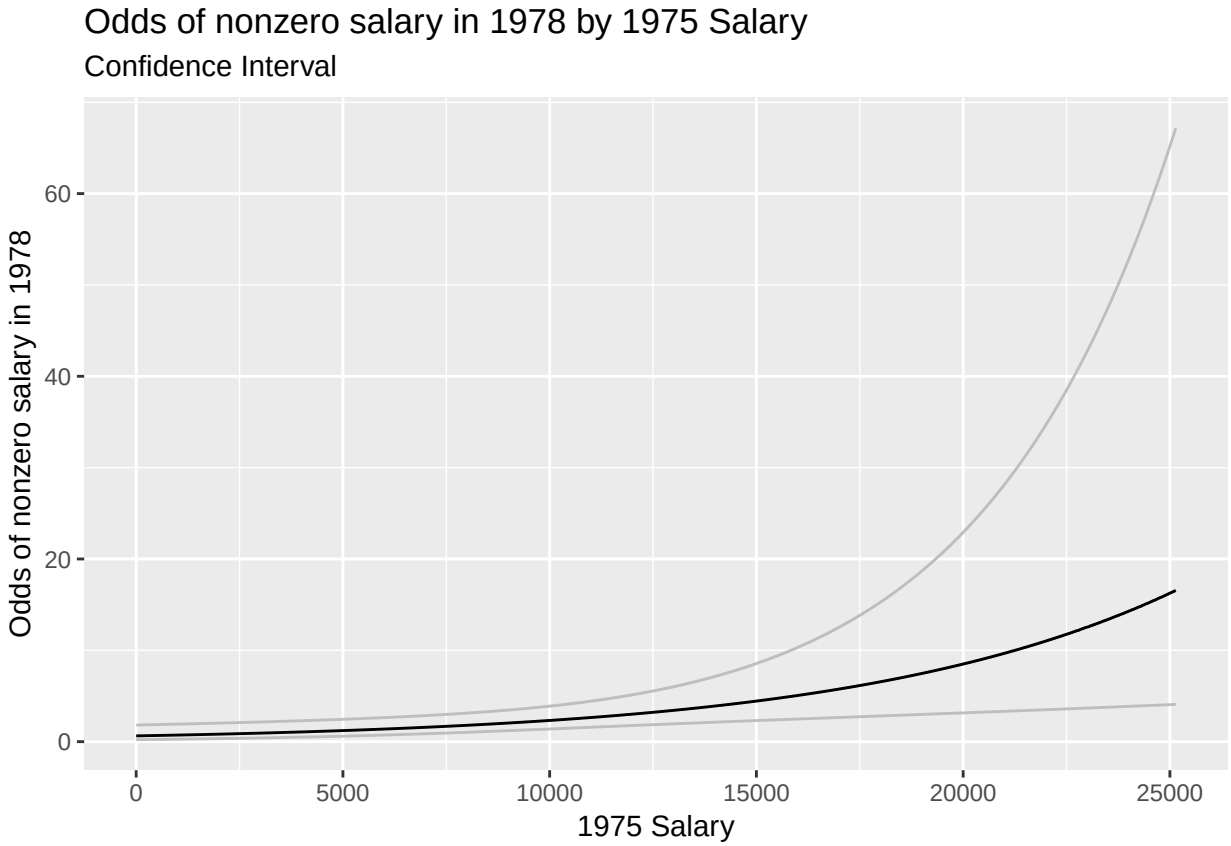
Age: Holding all else constant, for each 10 years an individual ages on average we estimate his salary to decrease by 0.72 (95% CI: 0.59, 0.89).

Married: Holding all else constant, for married individuals we estimate estimate the odds of nonzero wage in 1978 to increase by a factor of 1.05 more (95% CI: $0.65, 1.69). Given that this confidence interval includes 1, we are not confident that there is a meaningful effect of being married on odds of nonzero salary in 1978.

Black: Holding all else constant, for black individuals we estimate estimate the odds of nonzero wage in 1978 to decrease by a factor of 0.59 less (95% CI: $0.35, 1).

Hispanic: Holding all else constant, for black individuals we estimate estimate the odds of nonzero wage in 1978 to increase by a factor of 1.23 more (95% CI: 0.63, 2.59). Given that this confidence interval includes 1, we are not confident that there is a meaningful effect of hispanic ethnicity on odds of nonzero salary in 1978.

1975 Salary: Holding all else constant, for each $1,000 an individual made in 1975, on average we estimate estimate the odds of nonzero wage in 1978 to increase by a factor of 1.14 higher (95% CI:1.05, $1.25).



1974 Salary:

```
## Waiting for profiling to be done...
```

*Zero Salary: Holding all else constant, for individuals who participated in the treatment we estimate the odds of nonzero wage in 1978 to increase by a factor of 2.17 (95% CI: 1.15, 4.12).

*Nonzero Salary: Holding all else constant, for individuals who participated in the treatment we estimate the odds of nonzero wage in 1978 to decrease by a factor of 0.74 (95% CI: 0.34, 1.67). Given that this confidence interval includes 1, we are not confident that there is a meaningful effect of treatment on nonzero salary in 1978 for those with nonzero salaries in 1974.

```
# Create dummy dataset for charting
newval_74 <- data.frame(treat = as.factor(c(0,0,1,1)),
                        employed74 = c(0,1,0,1),
                        agec = 0,
                        married = 0,
                        black = 0,
                        hispan = 0,
                        educ.bin2 = 'Some HS +',
                        re75c = 0)

# Predict responses
predict <- predict.glm(final_log_fit, newval_74, interval = 'response', se.fit = TRUE)

# Create confidence interval
t <- 1.96 ## approx 95% CI
upr <- predict$fit + (t * predict$se.fit)
lwr <- predict$fit - (t * predict$se.fit)
```

```r
fit <- predict$fit

# Append predictions
newval_74 <- newval_74 %>%
    mutate(fit = exp(fit),
           lwr = exp(lwr),
           upr = exp(upr))

newval_74 %>%
    mutate(employed74 = ifelse(employed74 == 0, 'Zero 1974 Salary', 'Nonzero 1974 Salary')) %>%
    ggplot() +
    facet_grid(. ~ employed74) +
    geom_point(mapping = aes(x = treat, y = fit) ) +
    geom_point(mapping = aes(x = treat, y = lwr), shape = 1) +
    geom_point(mapping = aes(x = treat, y = upr), shape = 1) +
    ylab('Odds of nonzero salary in 1978') +
    xlab('Treament') +
    ggtitle('Interaction effect of nonzero 1974 salary on treatment') +
    labs(subtitle = 'Confidence Interval')
```

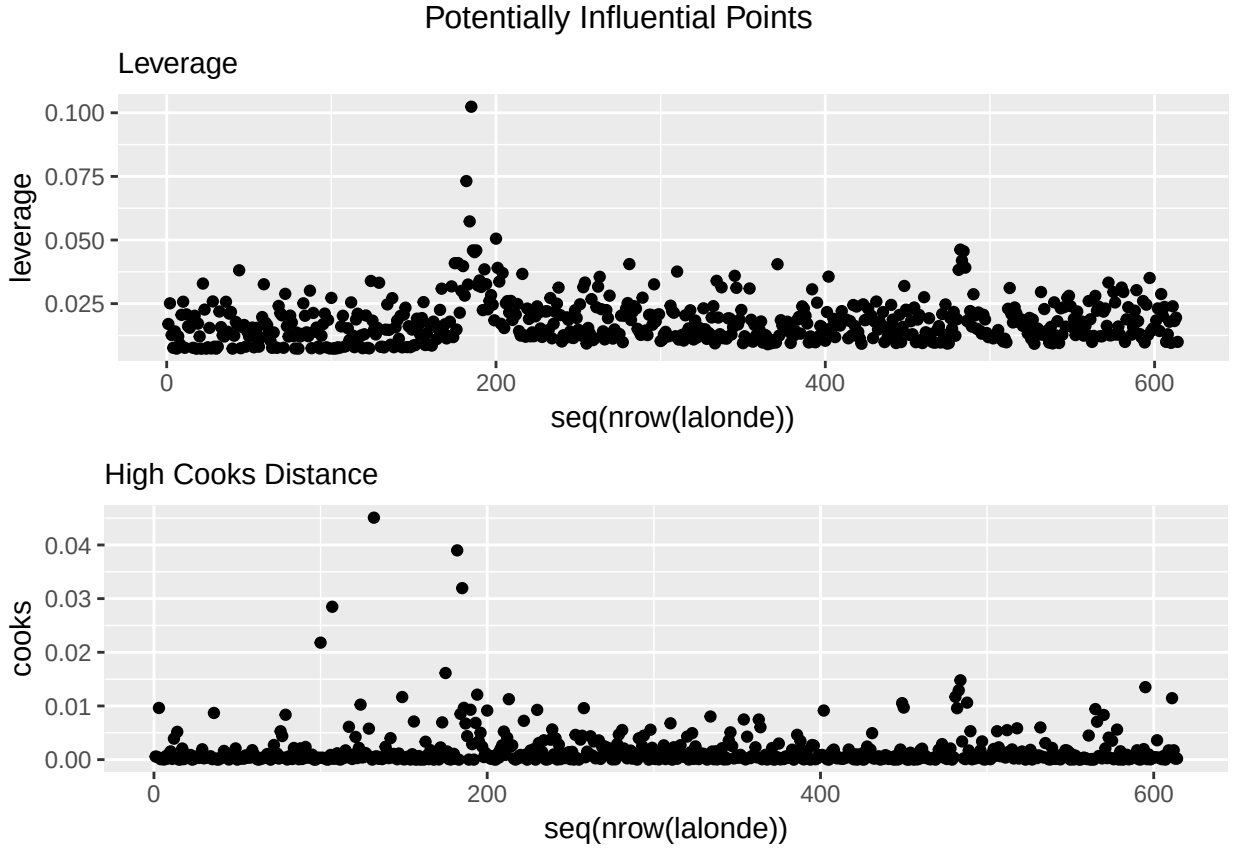
Appendix

Fig 1: Linear Interaction Plots Fig 2: Linear Residual Plots

Influential Points

Linear Model

Observations with high leverage or cooks distance in our final linear model are below:



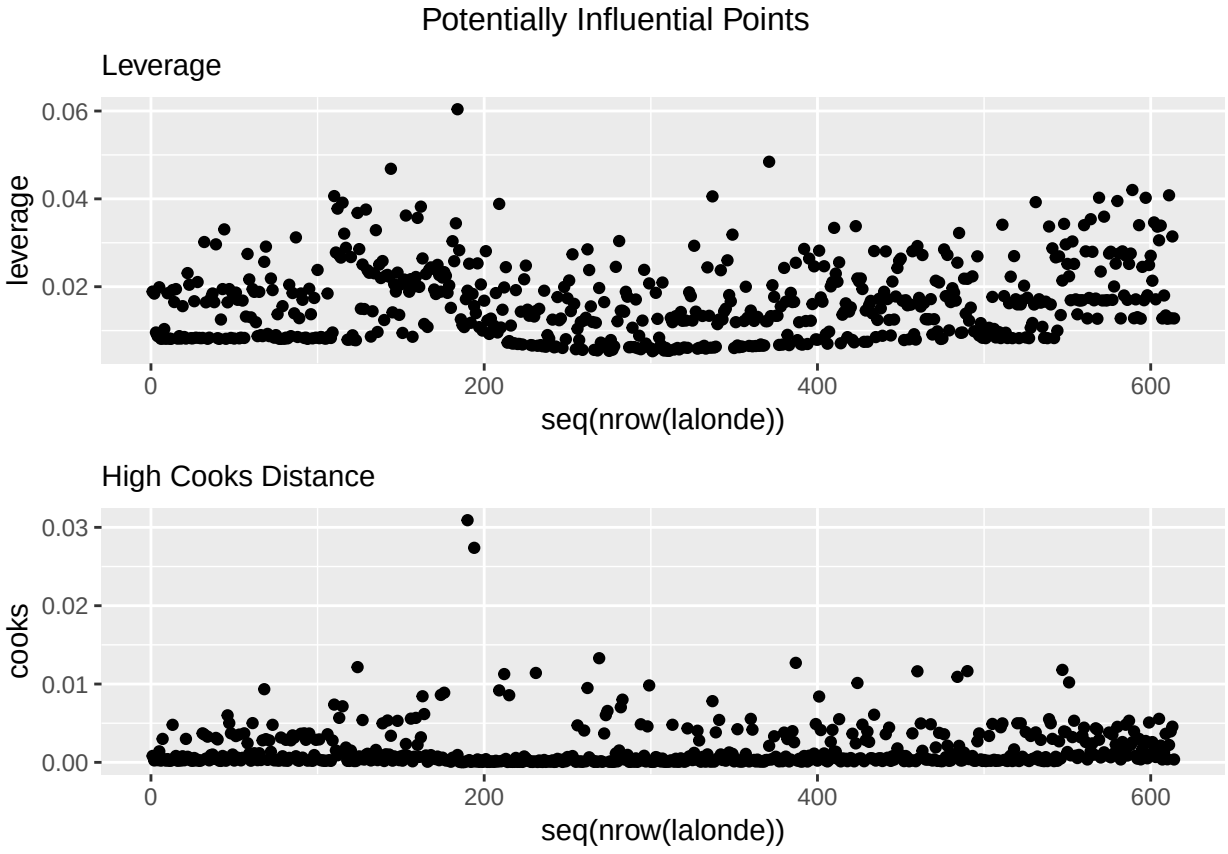

```
##        X treat age educ black hispan married nodegree       re74      re75
## 1 NSW100     1  31    9     0      1       0        1     0.0000     0.000
## 2 NSW107     1  27   13     1      0       0        0     0.0000     0.000
## 3 NSW132     1  28   11     1      0       0        1     0.0000  1284.079
## 4 NSW182     1  25   14     1      0       1        0 35040.0700 11536.570
## 5 NSW184     1  35    8     1      0       1        1 13732.0700 17976.150
## 6 NSW185     1  33   11     1      0       1        1 14660.7100 25142.240
## 7 PSID15     0  22   14     1      0       1        0   748.4399 11105.370
##        re78     re78c      re75c     re74c       agec employed78
## 1 26817.600 20024.766 -2184.9382 -4557.547  3.6368078          1
## 2 34099.280 27306.446 -2184.9382 -4557.547 -0.3631922          1
## 3 60307.930 53515.096  -900.8592 -4557.547  0.6368078          1
## 4 36646.950 29854.116  9351.6318 30482.523 -2.3631922          1
## 5  3786.628 -3006.206 15791.2118  9174.523  7.6368078          1
## 6  4181.942 -2610.892 22957.3018 10103.163  5.6368078          1
## 7 18208.550 11415.716  8920.4318 -3809.107 -5.3631922          1
##   employed74     educ.bin  educ.bin2       age2          age3    leverage
## 1          0      Some HS  Some HS + 13.2263711   48.10176982 0.027257833
## 2          0 More than HS  Some HS +  0.1319086   -0.04790816 0.020244859
## 3          0      Some HS  Some HS +  0.4055242    0.25824098 0.007797395
## 4          1 More than HS  Some HS +  5.5846773  -13.19766572 0.073164174
## 5          1   MS or less MS or less 58.3208336  445.38499829 0.057317557
```

```
## 6          1      Some HS  Some HS + 31.7736024  179.10169025 0.102376529
## 7          1 More than HS  Some HS + 28.7638304 -154.26595026 0.050528522
##         cooks
## 1 0.021798727
## 2 0.028482254
## 3 0.045086361
## 4 0.038990285
## 5 0.008516218
## 6 0.031957785
## 7 0.009140412
```

The influential points show that our model is not as accurate in its predictions for those who have high salaries in either 1974 or 1975. Because these are not the typical demographic to partake in a job training program, they are not of great interest for this research paper. Therefore we do not alter our model.

Logistic Model

Observations with high leverage or cooks distance in our final logistic model are below:




```
##        X treat age educ black hispan married nodegree     re74     re75
## 1 NSW184     1  35    8     1      0       1        1 13732.07 17976.15
## 2  PSID5     0  25    9     1      0       1        1 14829.69 13776.53
## 3  PSID9     0  38    9     0      1       1        1 16826.18 12029.18
##       re78     re78c     re75c     re74c      agec employed78 employed74
## 1 3786.628 -3006.206 15791.212  9174.523  7.636808          1          1
## 2    0.000 -6792.834 11591.592 10272.143 -2.363192          0          1
## 3    0.000 -6792.834  9844.242 12268.633 10.636808          0          1
##     educ.bin  educ.bin2       age2       age3   leverage        cooks
```

```
## 1 MS or less MS or less  58.320834  445.38500 0.06040330 0.0009428119
## 2    Some HS  Some HS +   5.584677  -13.19767 0.01909379 0.0309170281
## 3    Some HS  Some HS + 113.141681 1203.46631 0.01561386 0.0273889532
```

The influential points show that our model is not as accurate in its predictions for those who have high salaries in either 1974 or 1975. Because these are not the typical demographic to partake in a job training program, they are not of great interest for this research paper. Therefore we do not alter our model.

Any other limitations plots + outlier