Team Project 1

Allison Young and Anna Berman

10/25/2018

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.

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 logistic regression is summarized below:

```
##  treat        age             educ           black            hispan
##  0:429   Min.   :16.00   Min.   : 0.00   Min.   :0.0000   Min.   :0.0000
##  1:185   1st Qu.:20.00   1st Qu.: 9.00   1st Qu.:0.0000   1st Qu.:0.0000
##          Median :25.00   Median :11.00   Median :0.0000   Median :0.0000
##          Mean   :27.36   Mean   :10.27   Mean   :0.3958   Mean   :0.1173
##          3rd Qu.:32.00   3rd Qu.:12.00   3rd Qu.:1.0000   3rd Qu.:0.0000
##          Max.   :55.00   Max.   :18.00   Max.   :1.0000   Max.   :1.0000
##     married          nodegree          re74            re75
##  Min.   :0.0000   Min.   :0.0000   Min.   :    0   Min.   :    0.0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:    0   1st Qu.:    0.0
##  Median :0.0000   Median :1.0000   Median : 1042   Median :  601.5
##  Mean   :0.4153   Mean   :0.6303   Mean   : 4558   Mean   : 2184.9
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 7888   3rd Qu.: 3249.0
##  Max.   :1.0000   Max.   :1.0000   Max.   :35040   Max.   :25142.2
##       re78             re78c           re75c           re74c
##  Min.   :    0.0   Min.   :-6793   Min.   :-2185   Min.   :-4558
##  1st Qu.:  238.3   1st Qu.:-6555   1st Qu.:-2185   1st Qu.:-4558
##  Median : 4759.0   Median :-2034   Median :-1583   Median :-3515
##  Mean   : 6792.8   Mean   :    0   Mean   :    0   Mean   :    0
##  3rd Qu.:10893.6   3rd Qu.: 4101   3rd Qu.: 1064   3rd Qu.: 3331
##  Max.   :60307.9   Max.   :53515   Max.   :22957   Max.   :30483
##       agec           employed78      employed75      employed74
##  Min.   :-11.363   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
```

```
##  1st Qu.: -7.363   1st Qu.:1.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median : -2.363   Median :1.0000   Median :1.000   Median :1.0000
##  Mean   :  0.000   Mean   :0.7671   Mean   :0.601   Mean   :0.6042
##  3rd Qu.:  4.637   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   : 27.637   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##          educ.bin        educ.bin2
##  HS           :157   Some HS + :480
##  MS or less   :134   MS or less:134
##  Some HS      :253
##  More than HS: 70
##
##
```

Linear Regression

Exploratory Data Analysis

For concern of multicolinearity, 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 multicolinearity 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 (For further details on considered interaction effects ,see appendix). Ultimately we included a binned version of education seperating those with no high school education, some high school education, high school completion, and additional education beyond high school (see exploratory plots for incentive for variable alteration). 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 residuals 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
```

```
##                           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

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 receive 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 being married 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 Hispanic 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

Our findings suggest that participation in the examined job training results in increased salaries. Because this is a randomized control trial, we can say this is a casual effect. However the effect size may be small. Specifically we estimate that individuals who participated in job training to have average 1978 salaries increased by $1612.92, however this effect could be as smaller as $81.26 or as large as $3144.58.

Additionally, our findings suggest that 1978 salary is also mediated by level of education and previous salary. Specifically, education above high school is positively associated with 1978 salary and education level below 9th grade being negatively associated with 1978 salary. In other words, it is unclear whether having a high school diploma differs significantly from having some high school education when it comes to salary in 1978. However, it is clear that having less than a high school education results in a lower salary, and having more than a high school education results in a higher salary.

Simiarly, higher 1974 and 1975 salaries are both indepedently associated with increased 1978 salaries. Interestingly, the relationship betweetn 1974 and 1978 salary appears to be stronger than that between 1975 and 1978 salary. In other words, it appears the 1974 salary of individuals is more representative of earning potential in 1978 compared to 1975, and thus a stronger predictor factor of 1978 salary in our final model.

Contrastly, age, race, ethnicity and marital status may or may not have an effect on salary. This is evident because, when isolated, the cofidence intervals of the model coefficients include zero.

Limitations

Our model has an R-squared of 0.15. In other words, our model explains 15% of the variance 1978 salary. These results suggest that there are additional variables that may be stronger predictors of salary than are

included in our dataset. More research is needed to fully understand the relationship between job training programs and salary and the mediating variables in this relationship.

Additionally, our model appears to be less predictive for those with relatively high 1974 or 1975 salaries. This may be due to the lack of observations including high starting salaries. Additional research is needed to fully understand the relationship between salary and job trainings for those high above average salary before job training.

Logistic Regression

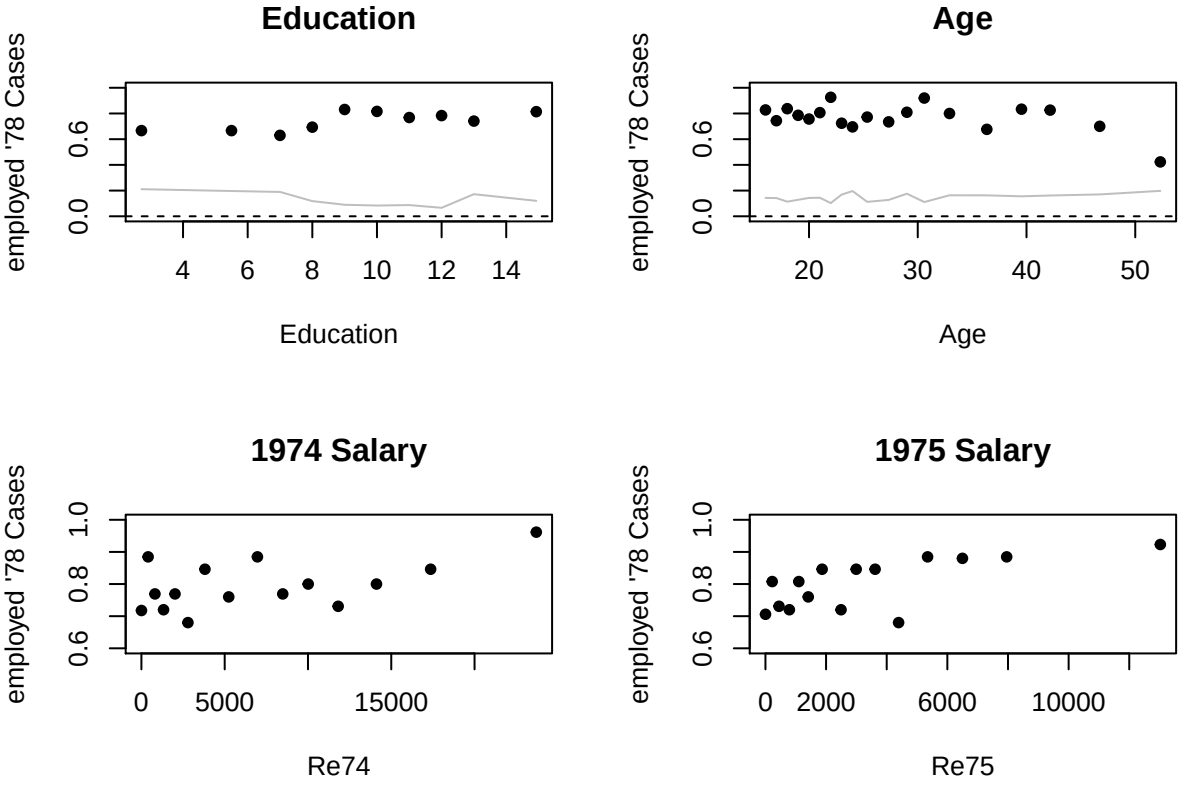
Exploratory Data Analysis

In terms of mulicollinearity, by the same reasoning as decribed in our linear regression summary, our only restriciton is a mutually exclusive 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.

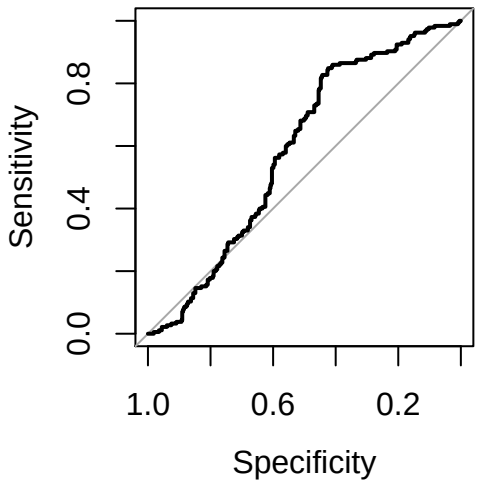
We attempted binning education in multiple ways, adding nonzero 1974 and 1975 salary variables. We also examined potential interactions between treatment and previous salaries as well as interactions between treatment and level of education (see appendix for examination of interaction effects). Ultimately we included a binned version of education seperating those with at least some high school education or more and those with no high school education (see exploratory plots for incentive for variable alteration). The only interaction that resulted in a significant change in deviance test was the addition of an interaction effect between nonzero 1974 salary and treatement.

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 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
```

```
## [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

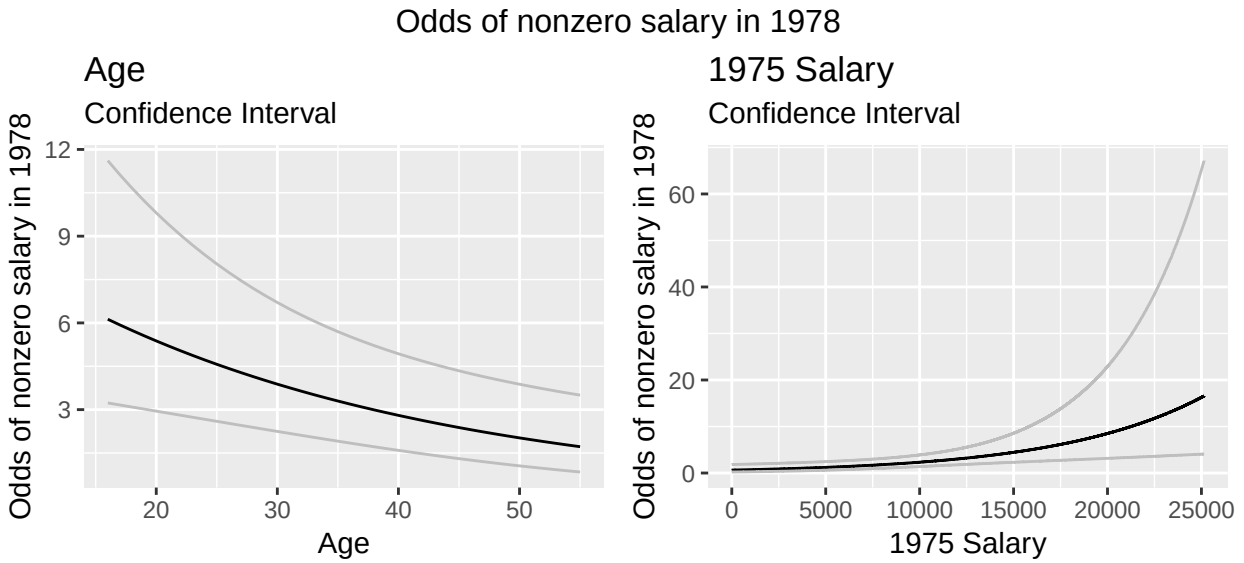
Our model has an AUC of 0.59. Using the suggested threshold of 0.82, our model has a sensitivity of 0.391 and a specificity of 0.175. In other words, our model correctly predicts 39.1% of nonzero wage earners and 82.5% of zero wage earners.

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 salary 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 the odds of nonzero salary 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 odds of nonzero salary in 1978 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.

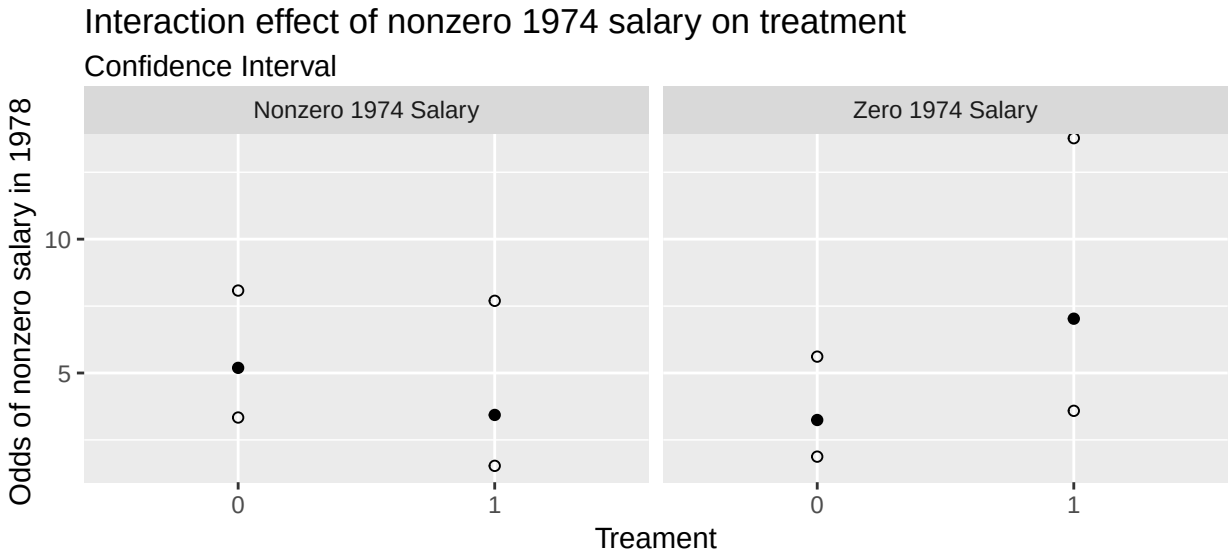
Black: Holding all else constant, for Black individuals we estimate estimate the odds of nonzero wage in 1978 to decrease by a factor of 0.59 less (95% CI: 0.35, 1).

Hispanic: Holding all else constant, for Hispanic 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 (Zero vs. Nonzero):

- 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.



Discussion

According to our findings, participating in the training program has a positive effect on employment status for individuals who were unemployed in 1974, increasing the odds of a non-zero salary in 1978 by an estimated factor of 2.17. However, we can not say with certainty whether the impact is the same for those who were employed in 1974.

For those employed in 1974 our estimates indicate that the job training program may actually have a negative impact, decreasing odds of employment in 1978 by an estimated factor of 0.74. There appears to be a negative effect on the odds of employment for those who were employed in 1974, we are not able to say with confidence that this effect is negative. Moreover, we are unable to say an interaction between treatment and employed status in 1974 exists. The confidence intervals overlap, and as such, the true odds ratio may be the same for both groups.

Odds of employment in 1978 are also influenced by level of education, age, being black. Speicifically, level of education not including any high school, increased age, and blackness are all associated with decreased odds of nonzero wage in 1978. In other words, increased age, an elementary or middle-school level education, or being black are all associated with increased odds of zero wages. Contrastly, higher 1975 salary is assocaited with increased odds of nonzero wage in 1978.

Additionally, being Hispanic and being married may or may not be associated with odds of employment in 1978. When all other factors are held constant, each of these factors has an odds ratio confidence interval that includes 1.0.

Limitations

There are several limitations to our model that should be taken into consideration when interpreting results. First, while the confidence interval for the odds ratio of employment in 1978 for an individual in a job training program doesn't include 1.0, it is very close to 1.0. Therefore, while a positive effect is likely, it may not be large.

Second, our resulting model has a low area under the curve (AUC) of 0.59. Therefore, an AUC of 0.59 is on the lower end of the possible range of .5-1.0. Specific to our model, there is a sizable difference in the sensitivity (.820) and specificity '(.424) of our model. Meaning, while our model is good at predicting individuals likely to be unemployed in 1978, it is not as good at capturing the breadth of individuals who are truly employed.

Conclusion

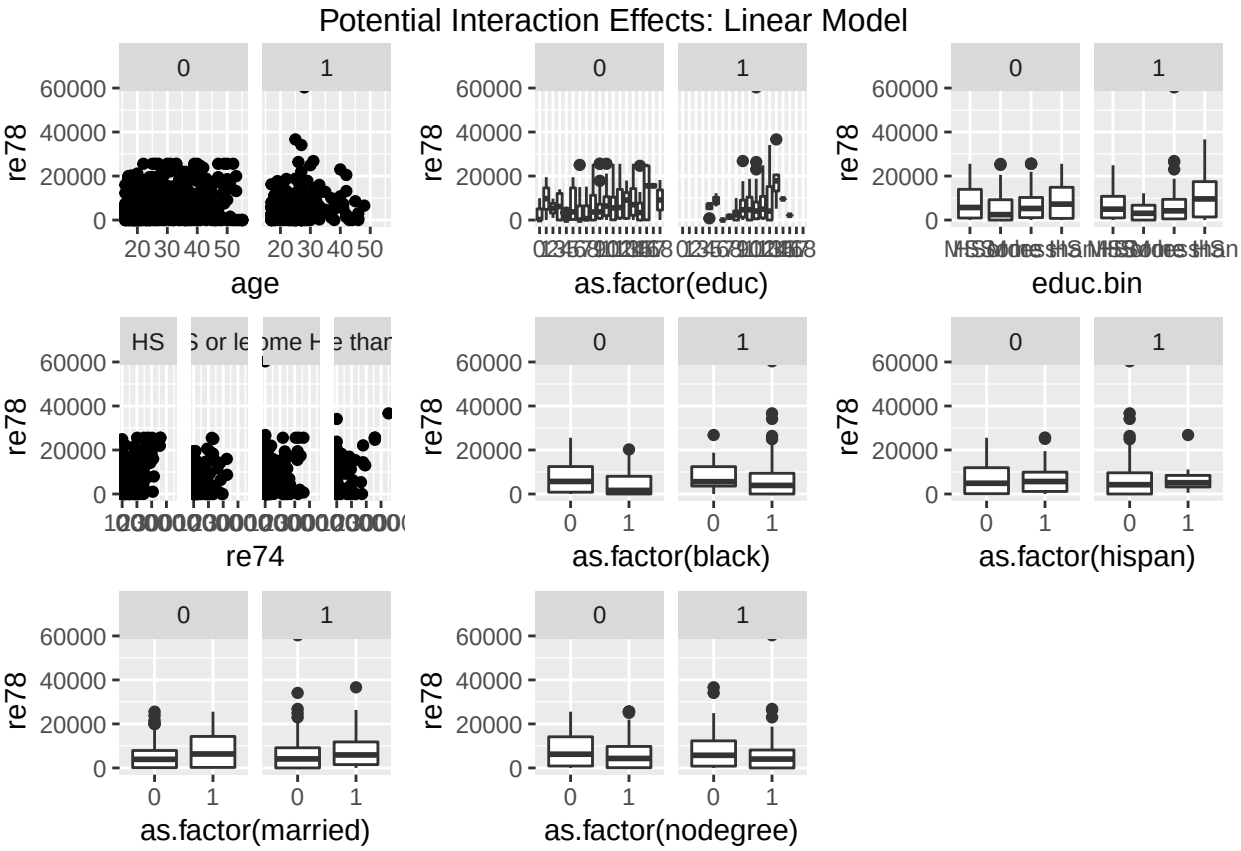
Based on our models, there is good evidence that job training positively influences the odds of employment and the salaries of individuals who complete the program. However, neither our linear model nor our logistic model provide confidence in large increases in odds of employment nor salaries, as they both have large confidence intervals.

Appendix

Interaction Exploration

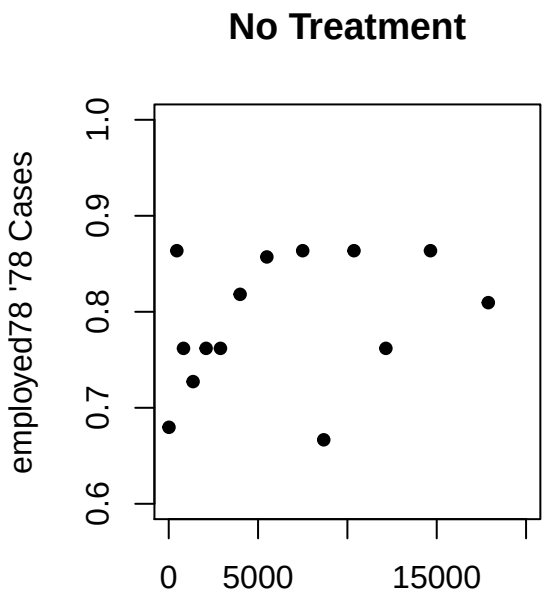
Below is an exploration of potential interaction effects between treatment and other predictor variables in our dataset. Although some exploratory analysis seemed promising, no addition of interaction terms meaningfully increased our models performance. Therefore our final model did not include any interaction effects.

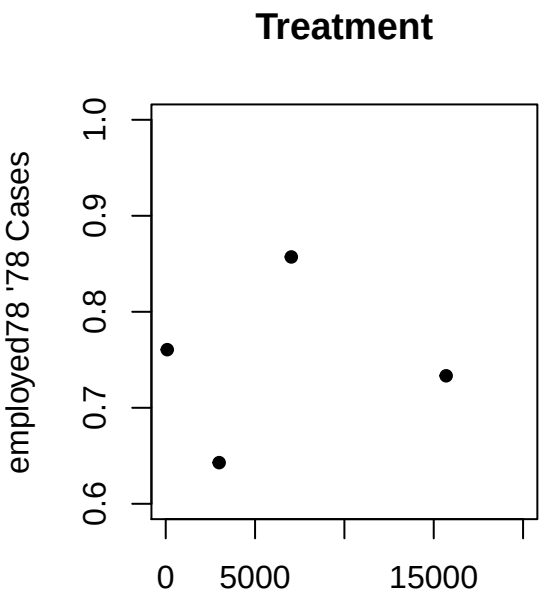
Linear Model

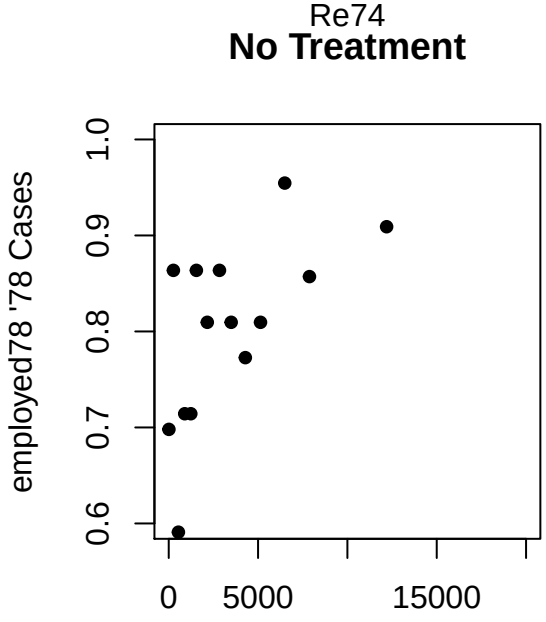


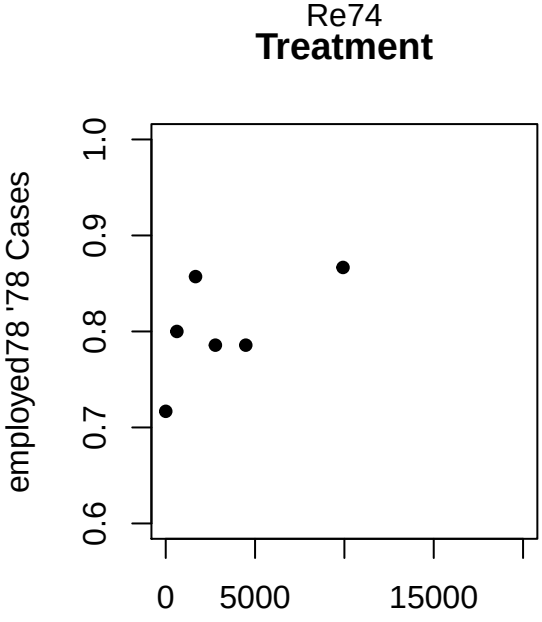
Logistic Model

Below is an exploration of potential interaction effects between treatment and other predictor variables in our dataset. The only interaction effect that meaningfully improved prediction for our model was the addition an interaction between a nonzero employment in 1974 and treatment.









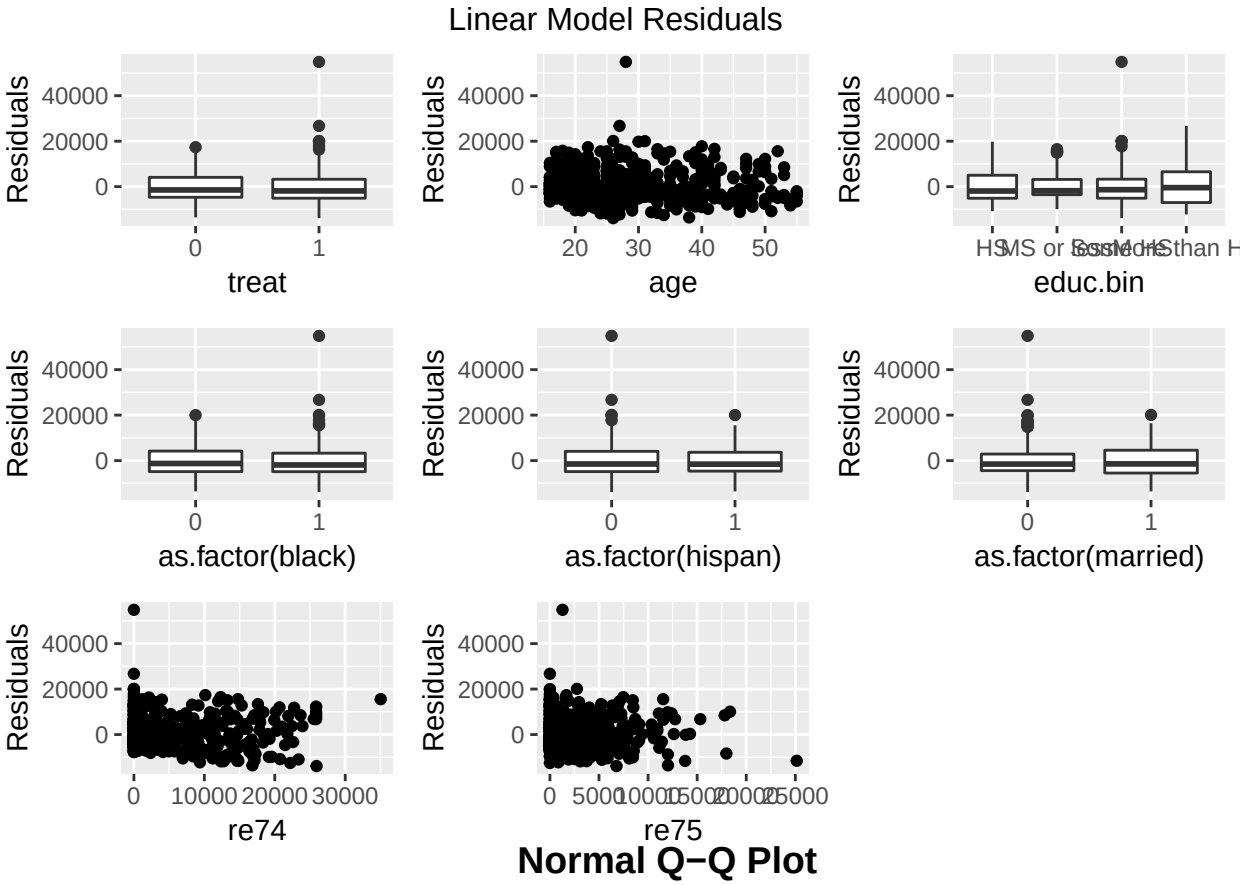

```
##              Unemployed 75 Employed 75
## No Treatment     0.6865672   0.8101695
## Treatment        0.7207207   0.8108108
```

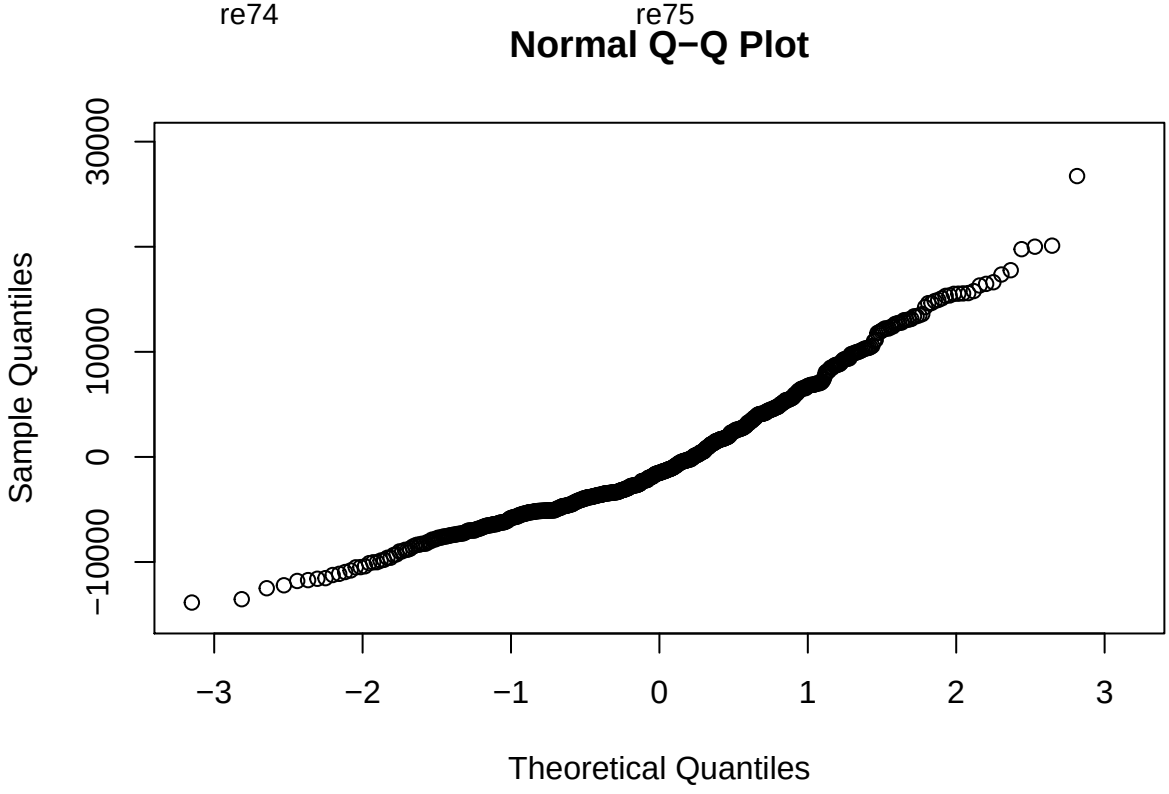
```
##              Unemployed 74 Employed 74
## No Treatment     0.6696429   0.8075710
## Treatment        0.7633588   0.7407407
```

```
##              Some HS + MS or less
## No Treatment 0.8049536  0.6698113
## Treatment    0.7707006  0.6785714
```

Residuals

Linear Model





Logistic Model

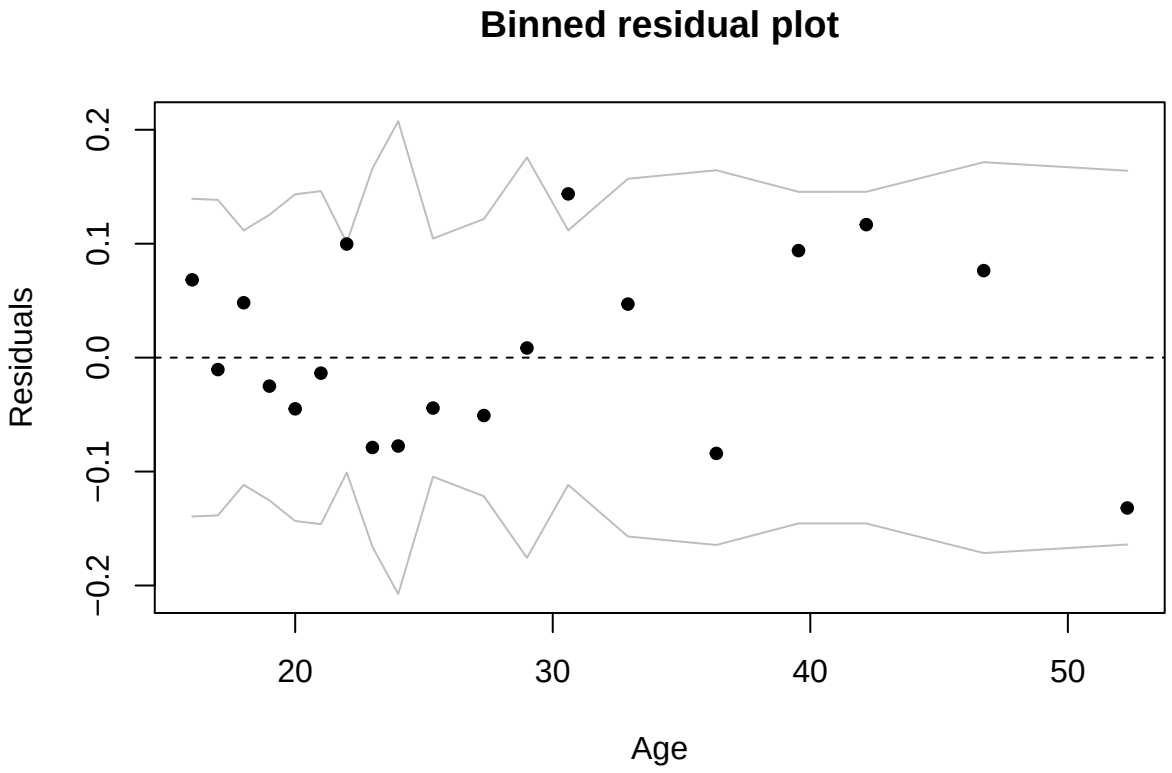
```
##    Some HS +   MS or less
## 1.120612e-09 1.286734e-09
```

```
##            0            1
## 1.230703e-09 1.044137e-09
```

```
##            0            1
## 1.134503e-09 1.325221e-09
```

```
##            0            1
## 3.727226e-10 2.260819e-09
```

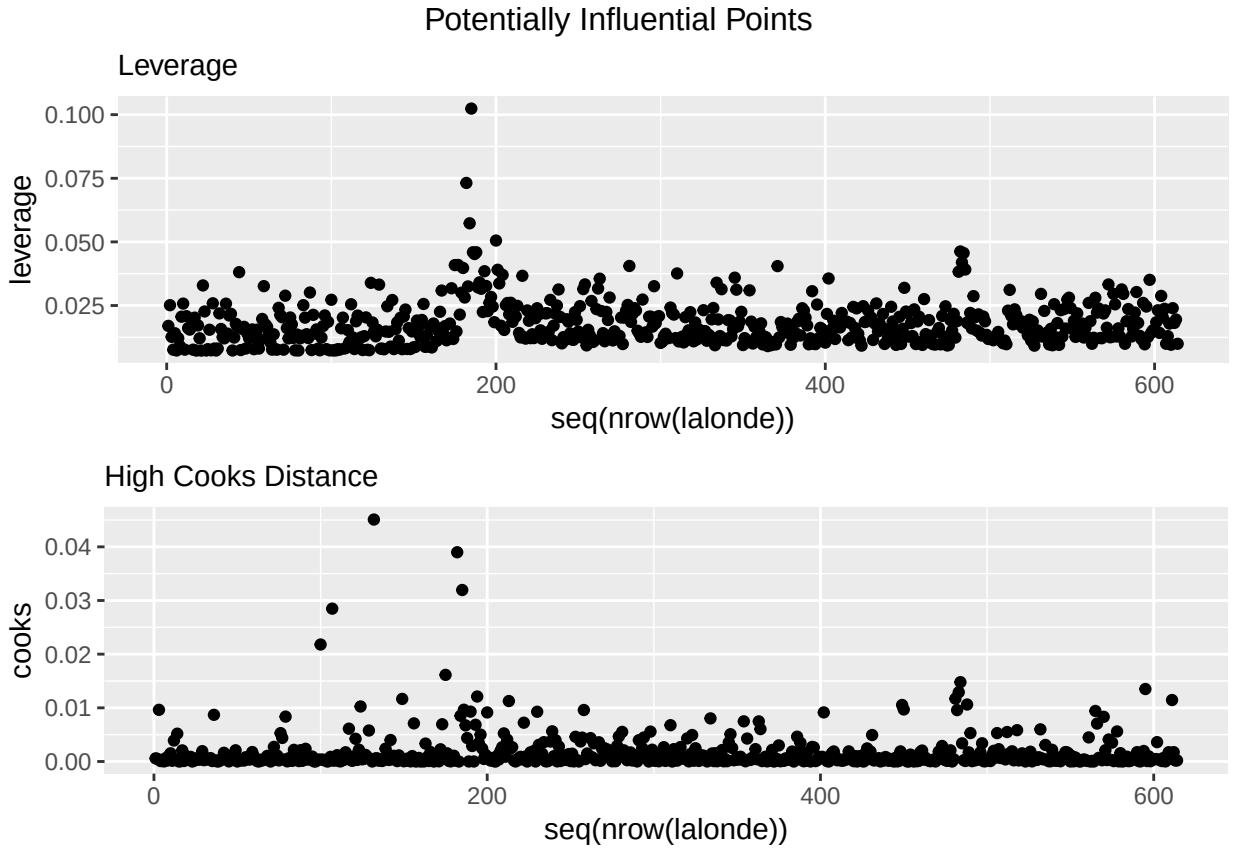
```
##            0            1
## 1.668805e-10 1.805295e-09
```



Influential Points

Linear Model

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



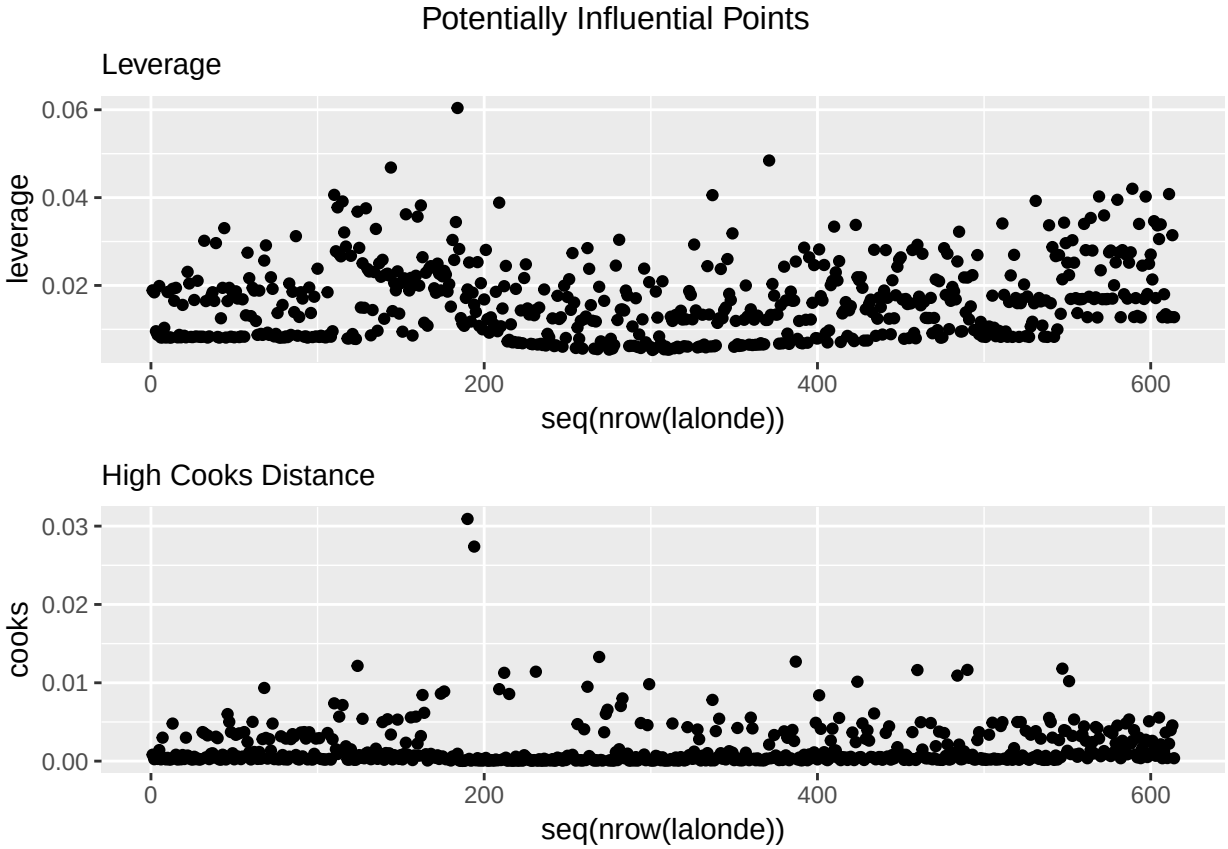

```
##        X treat age educ black hispan married nodegree       re74      re75
## 1 NSW100     1  31    9     0      1       0        1     0.0000     0.000
## 2 NSW107     1  27   13     1      0       0        0     0.0000     0.000
## 3 NSW132     1  28   11     1      0       0        1     0.0000  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
##   employed75 employed74     educ.bin  educ.bin2       age2          age3
## 1          0          0      Some HS  Some HS + 13.2263711   48.10176982
## 2          0          0 More than HS  Some HS +  0.1319086   -0.04790816
## 3          1          0      Some HS  Some HS +  0.4055242    0.25824098
## 4          1          1 More than HS  Some HS +  5.5846773  -13.19766572
## 5          1          1   MS or less MS or less 58.3208336  445.38499829
## 6          1          1      Some HS  Some HS + 31.7736024  179.10169025
## 7          1          1 More than HS  Some HS + 28.7638304 -154.26595026
##      leverage       cooks
## 1 0.027257833 0.021798727
## 2 0.020244859 0.028482254
```

```
## 3 0.007797395 0.045086361
## 4 0.073164174 0.038990285
## 5 0.057317557 0.008516218
## 6 0.102376529 0.031957785
## 7 0.050528522 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
## 1   NSW110     1  26   10     1      0       1        1  2027.9990
## 2   NSW124     1  27   13     0      0       1        0  9381.5660
## 3   NSW144     1  46    8     1      0       0        1  3165.6580
## 4   NSW184     1  35    8     1      0       1        1 13732.0700
## 5    PSID5     0  25    9     1      0       1        1 14829.6900
## 6    PSID9     0  38    9     0      1       1        1 16826.1800
## 7   PSID84     0  37   11     0      1       0        1   615.2098
## 8  PSID152     0  52    0     0      1       1        1   773.9104
## 9  PSID186     0  53   10     0      1       0        1  7878.2120
## 10 PSID202     0  20    9     0      1       1        1     0.0000
## 11 PSID384     0  31    4     0      1       0        1     0.0000
## 12 PSID404     0  55    7     0      0       0        1     0.0000
```

```
## 13 PSID412     0  53   12     1      0       0        0     0.0000
## 14 PSID426     0  24    1     0      1       1        1     0.0000
##          re75      re78     re78c      re75c     re74c       agec
## 1      0.0000     0.000 -6792.834 -2184.9382 -2529.548 -1.3631922
## 2    853.7225     0.000 -6792.834 -1331.2157  4824.019 -0.3631922
## 3   2594.7230     0.000 -6792.834   409.7848 -1391.889 18.6368078
## 4  17976.1500  3786.628 -3006.206 15791.2118  9174.523  7.6368078
## 5  13776.5300     0.000 -6792.834 11591.5918 10272.143 -2.3631922
## 6  12029.1800     0.000 -6792.834  9844.2418 12268.633 10.6368078
## 7   4713.9190     0.000 -6792.834  2528.9808 -3942.337  9.6368078
## 8   2506.4520     0.000 -6792.834   321.5138 -3783.636 24.6368078
## 9   1489.5480 13170.980  6378.146  -695.3902  3320.665 25.6368078
## 10  1283.6610     0.000 -6792.834  -901.2772 -4557.547 -7.3631922
## 11     0.0000  1161.493 -5631.341 -2184.9382 -4557.547  3.6368078
## 12     0.0000     0.000 -6792.834 -2184.9382 -4557.547 27.6368078
## 13     0.0000     0.000 -6792.834 -2184.9382 -4557.547 25.6368078
## 14     0.0000 19464.610 12671.776 -2184.9382 -4557.547 -3.3631922
##    employed78 employed75 employed74     educ.bin  educ.bin2        age2
## 1           0          0          1      Some HS  Some HS +   1.8582929
## 2           0          1          1 More than HS  Some HS +   0.1319086
## 3           0          1          1   MS or less MS or less 347.3306056
## 4           1          1          1   MS or less MS or less  58.3208336
## 5           0          1          1      Some HS  Some HS +   5.5846773
## 6           0          1          1      Some HS  Some HS + 113.1416805
## 7           0          1          1      Some HS  Some HS +  92.8680649
## 8           0          1          1   MS or less MS or less 606.9722994
## 9           1          1          1      Some HS  Some HS + 657.2459151
## 10          0          1          0      Some HS  Some HS +  54.2165991
## 11          1          0          0   MS or less MS or less  13.2263711
## 12          0          0          0   MS or less MS or less 763.7931463
## 13          0          0          0           HS  Some HS + 657.2459151
## 14          1          0          0   MS or less MS or less  11.3110617
##             age3   leverage        cooks
## 1  -2.533210e+00 0.04062583 0.0073756699
## 2  -4.790816e-02 0.03679869 0.0121590942
## 3   6.473134e+03 0.04684611 0.0033930294
## 4   4.453850e+02 0.06040330 0.0009428119
## 5  -1.319767e+01 0.01909379 0.0309170281
## 6   1.203466e+03 0.01561386 0.0273889532
## 7   8.949517e+02 0.01969965 0.0132959816
## 8   1.495386e+04 0.04056232 0.0078171676
## 9   1.684969e+04 0.04844849 0.0021112125
## 10 -3.992072e+02 0.02544092 0.0127010077
## 11  4.810177e+01 0.04025713 0.0028860602
## 12  2.110880e+04 0.04201691 0.0025731029
## 13  1.684969e+04 0.04024095 0.0027277268
## 14 -3.804127e+01 0.04081239 0.0022232284
```

We see that many of influential points are those who have nonzero salary in 1974, but who have a zero salary in 1978. These are, in fact, outliers in our observation as they are people who seem to be meaningfully financially worse off in 1978 than in 1974. These cases are potentially influential in our model, but are scientifically relevant cases to include in our observations. Because we cannot justify removing these observations from our dataset, we leave our model unaltered.

Additionally, there are several outliers including relatively high 1978 salaries. Again, although this observation

is not the norm, we include this observation in our model as it is an important edge case.