MEMORANDUM

Should we target the job skills training program at individuals with or without high school degrees?

To: Alexis Diamond, Head of Program Management

From: Tomer Eldor, Analytics Department

Date: February 25, 2017

Subject: Running program for high-school graduates

EXECUTIVE SUMMARY

The program should be targeted towards individuals **with** a high school degree.

We analyzed results from an observational study taken at 1978, by Lalonde, through three different lenses: multiple (multivariate) linear regression, random forests, and Fisher's Exact Test. All three methods suggested that the program was indeed beneficial, and showed clearly that the program benefitted people with a high school degree far more than it benefitted people without a high school degree.

source: http://theknowledgeengineers.com/wp-content/uploads/2015/05/Digital-Training.jpg

Further Conclusions

We recommend running the program solely for participants with degree.

However, since the effect was small but still positive for people without a high-school degree, we suggest that in case that the program still has remaining funds left after covering all people with a degree, they could start accepting people with no degree. However, the program should be adapted to this segment, since it apparently is not highly effective for people lacking a high school degree. Alternatively, if the company has a better working program with a more significant positive impact than this impact on people without a degree, the company should simply reallocate these extra funds for that other program.

METHODS

We devised 3 regression tests: multiple (multivariate) linear regression, random forests, and Fisher's Exact Test.

1. LINEAR MULTIPLE REGRESSION.

Our multiple linear regression suggested a total average treatment effect of $1605.5 (for any individual, with no respect to having a degree or not).
That was done using a linear multiple regression upon all the other predictors (other than re78). The treatment effect is represented as the coefficient of the "treatment" predictor in the multivariate regression.

Assuming we set our significance threshold to be 5%, these results suggest that the treatment is statistically significant. The probability of this effect (or a more extreme one) occurring by chance is 1.25% (Pr(>|t|)=0.0125). However, if another statistician would choose a conservative alpha value of 1%, these results wouldn't pass as statistically significant (but wouldn't be too far from passing...).

In general, except for the huge negative effect that being black has on projected incomes (-2049 lower earnings), the treatment seems to have the **largest** coefficient of all binary predictors. Similarly, it has the lowest p-value (probability of occurring by chance). Therefore, we can say that between any other parameter measured in this experiment, treatment seems to have the

highest positive effect on projected earnings. The treatment is relatively more indicative than any other parameter, except for being black.
Short technical details of regression:

```
Call:
lm(formula = re78 ~ treat + educ + re75 + age + black + hisp +
   nodegr + u74 + u75 + married, data = lalonde)

Residuals:
  Min     1Q Median     3Q    Max
-9113  -4456  -1609   3048  53306

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  1285.0634  3448.6176   0.373   0.7096
treat       1605.5218   640.1676   2.508   0.0125 *
educ         371.7119   228.1403   1.629   0.1040
re75           0.1202     0.1323   0.908   0.3643
age           53.4331    45.8575   1.165   0.2446
black      -2049.8534  1174.9980  -1.745   0.0818 .
hisp         230.1306  1560.0865   0.148   0.8828
nodegr      -200.5898   998.3066  -0.201   0.8408
u74          622.3750  1059.5222   0.587   0.5572
u75         -841.9766  1012.9377  -0.831   0.4063
married     -155.1317   883.2236  -0.176   0.8607
```

The estimated confidence intervals for the coefficient estimates of each variable (Betas), of 95%, where from 347 to 2863.

Lower bound, upper bound confidence intervals:

```
> confint(lm2,level = 0.95)
            2.5 %      97.5 %
treat       347.3076270 2863.7360050
```

The confidence interval show that the treatment effect is significant, since it doesn't pass through 0. However, it shows us that we are highly uncertain by how much our treatment actually affects the results: the true average treatment effect could range between $347 to $2863. Meaning, our program might increase earnings in 1978, in average, by anything from $347 to $2863. An increase of $347 is **small** not suggesting that we should definitely invest resources into the program. But on the other hand, the opposite side - an increase of $2863, is highly impactful.

This regression does not allow for different effects for different individuals.

It only represents the **average treatment effect** for all individuals pulled together. If we would like to know how the effect differs for different individuals, we would need to perform some sort of stratification, blocking or division in our data, or divide the data into groups and perform our regressions separately upon each group, to get a group-specific regression and treatment effect. Theoretically, it could also occur in a multivariate regression, but rarely, and only if we would include *every possible interaction effect* in our formula.

Did the Linear regression suggest treatment effects were higher for individuals with (or without) a high school degree?

The treatment effects are estimated to be higher for individuals with degree than for people without degree, as observed by our multiple regression results.

To check how treatment effects differ for individuals with (or without) a high school degree, I split the dataset into two according to these, fitted a regression and checked the treatment effect.

Results:

For only people with no degree, the observed treatment effect was $1052 (with a P value of 0.13, meaning probability of occurring by chance of 13%). For people with degree, the observed treatment effect was $2283 (with a p value of 0.14, or probability of occurring by chance, of 14%). Notice that both p values were larger than 5%, showing insignificant results.

Note that if we perform confidence intervals tests, we see that not only both pass through 0, therefore the treatment effect for each of those groups could be in fact **null**, but also - the confidence intervals between these groups have a large overlap, therefore there might *not actually be* a significant difference:

Confidence intervals for people with No Degree:

treat	-333.58	2437.84

Confidence intervals for people with Degree:

treat	-810.05	5376.20

Therefore, we see that there is a big overlap between -810.05 and 2437.84. So, while that the treatment effect is different **on average** between groups, the true mean effect might actually be the same, or even smaller for people with a degree than people without a degree. Additionally, the confidence interval for people with degree is much wider, indicating that we have a less certainty about the treatment effect for people with degree, which decreases our certainty of results.

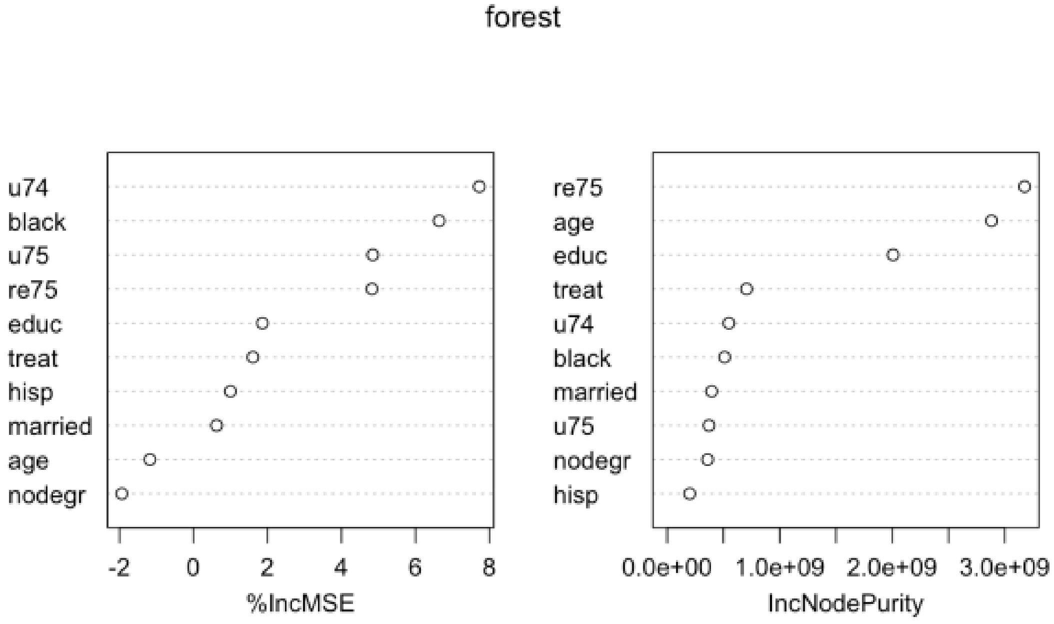
However, we must note that this method might be problematic because we are not performing proper matching and might have confounded results. Some variables might be too correlated, for example “black” and “no degree”, and thus by measuring the differences between “no degree” we are also confounded by the effects of being ‘black’.

2. RANDOM FOREST

We ran a random forest with the same variables as inputs as in the linear regression.
The results of the variable plot were as follows:

a. variable importance plot

```
                 %IncMSE IncNodePurity
          treat    2.3750166    304463569
          educ      1.9735474    494588321
          re75      2.8464480    993819784
          age      0.3813223    833077644
          black    6.9454653    215016847
          hisp     0.2241125    15786681
          nodegr   0.0000000            0
          u74      2.4920602    162972285
          u75      -0.5035656    101486304
married -2.0276482    105029325
```



We see the treatment effect for the all treated units is $**1273**.414. Meaning, the treated units are estimated to have gained $**1273**.414 on average by the program.

Both results support the that the treatment is significantly beneficial, although the random forest results predicts a bit smaller gain (treatment effect): $**1273** rather than $1605 predicted by the multiple linear regression.

Random Forests are much more flexible methods, and therefore could be that they are more accurately predicting since they are nonlinear and more accurately able to reflect the relations in our specific dataset. However, this is an observational dataset from a past program, so it might be not perfectly generalizable to our case. Therefore, a linear regression might have given us a more generalizable model with smaller variance (variability upon dataset), but potentially somewhat biased.

Similarly, our random forest does NOT produce different effects for different individuals. It predicts an average treatment effect for all units.

However, we can use the forest function to predict seperately for the 2 groups.

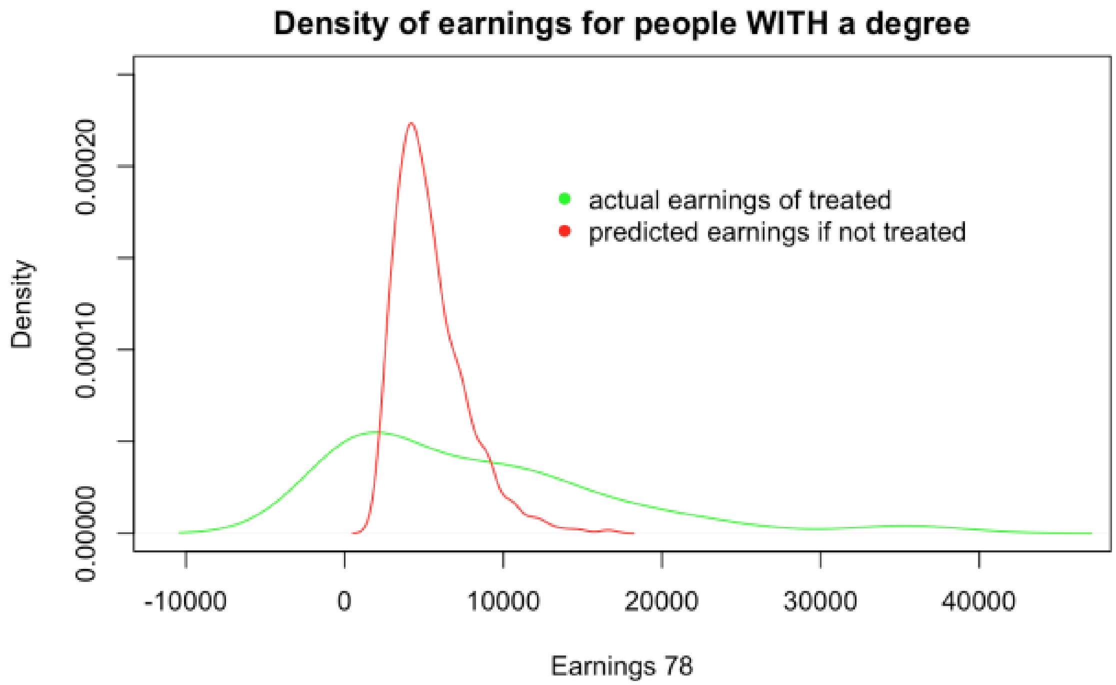
We performed this separate procedure and it showed a higher treatment effect for the people with a degree.

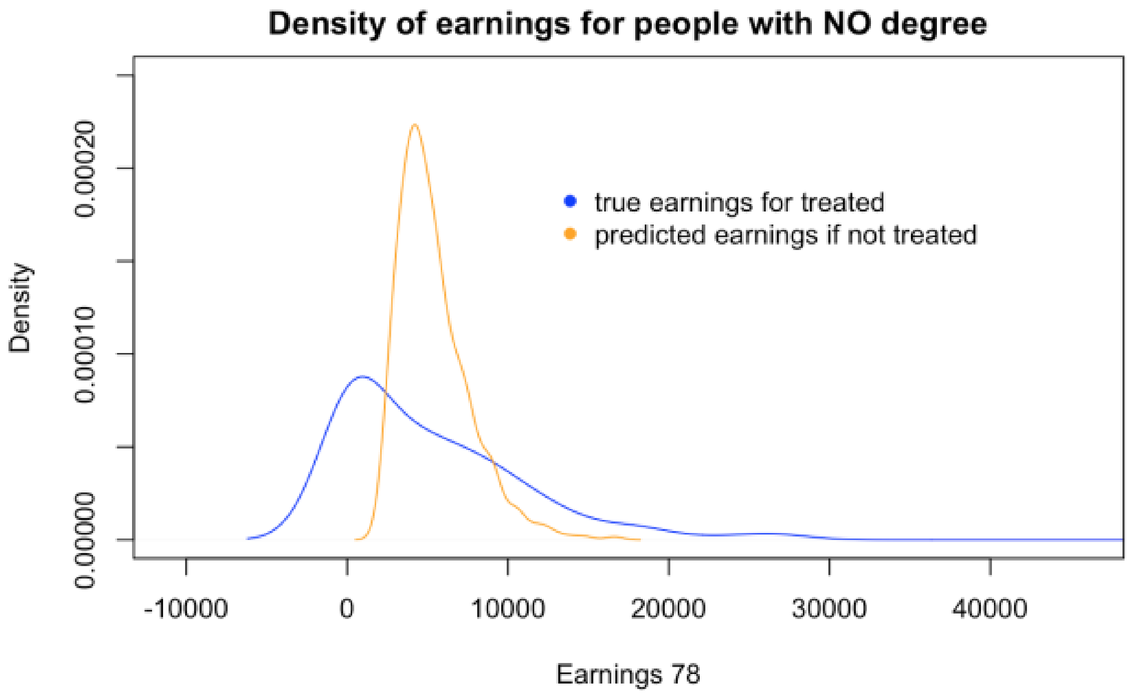
We then compared the potential outcomes for the people WITH or WITHOUT a DEGREE who were treated (in the program). This means we estimated how much would they have gained from whether they haven't been given the treatment.

The treatment effect for treated people WITH DEGREE is: $2647.82

So, the treatment effect for treated people WITH NO DEGREE is: $250.76 This is MUCH lower, almost miniscule, compared to the treatment effect for treated people WITH DEGREE, which was: $2647.82. Meaning, people with degree gained from the treatment, on average, **$2397.06** more than people without degree gained from the treatment. ($2647.82 vs $250.76).

The following 2 plots show the density distribution of the earnings for people **with or without a degree.**


Density of earnings for people WITH a degree
● actual earnings of treated
● predicted earnings if not treated
Density
0.00000
0.00010
0.00020
-10000
0
10000
20000
30000
40000
Earnings 78


Density of earnings for people with NO degree
● true earnings for treated
● predicted earnings if not treated
Density
0.00000
0.00010
0.00020
-10000
0
10000
20000
30000
40000
Earnings 78

3. IMPLEMENT A FISHER EXACT TEST (FET) USING THE SHARP NULL HYPOTHESIS OF NO TREATMENT EFFECT FOR ANY UNIT.

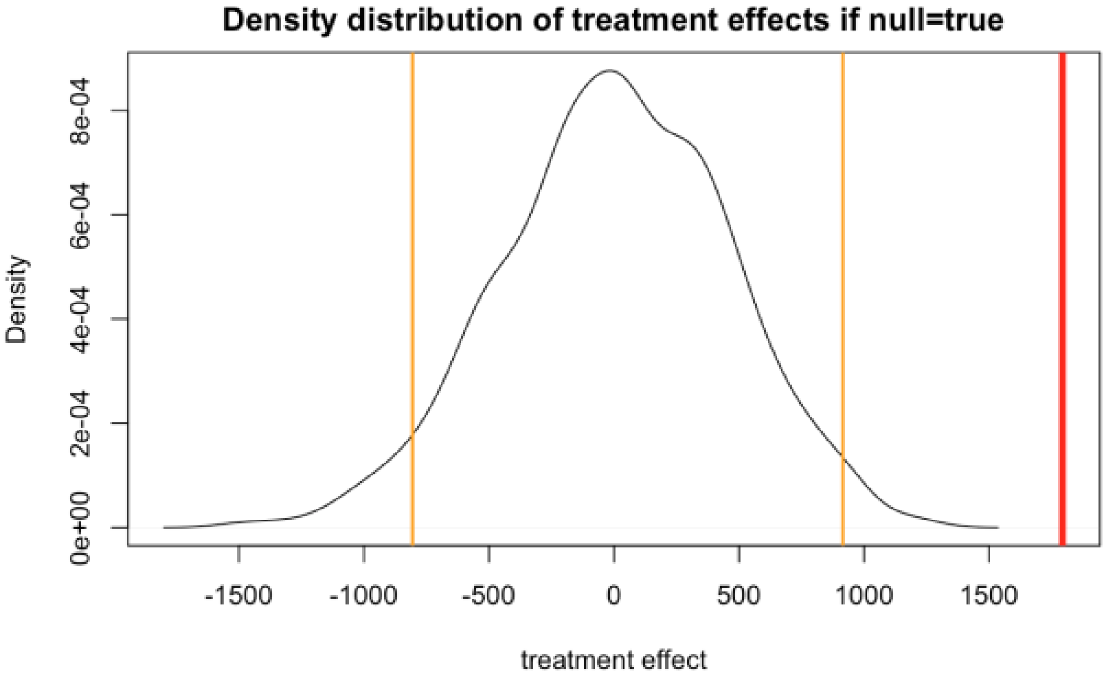
First, we define the test statistic as the mean of earnings for all treated units minus mean of all units in the control (**y(1) - y(0)**).

The original observed treatment effect between the original groups is: $1794.343.

We will define the sharp null hypothesis as "treatment effect = 0"" of absolutely no effect of treatment: (Yi(0) = Yi(1) for all units).

We simulated many hypothetical random assignments of the units to control and treatment groups, and check the treatment effects resulting from them. Then we check our how our original observed treatment effect stand in relation to these null hypothetical effects.

In brief, our results showed significance for the general data:

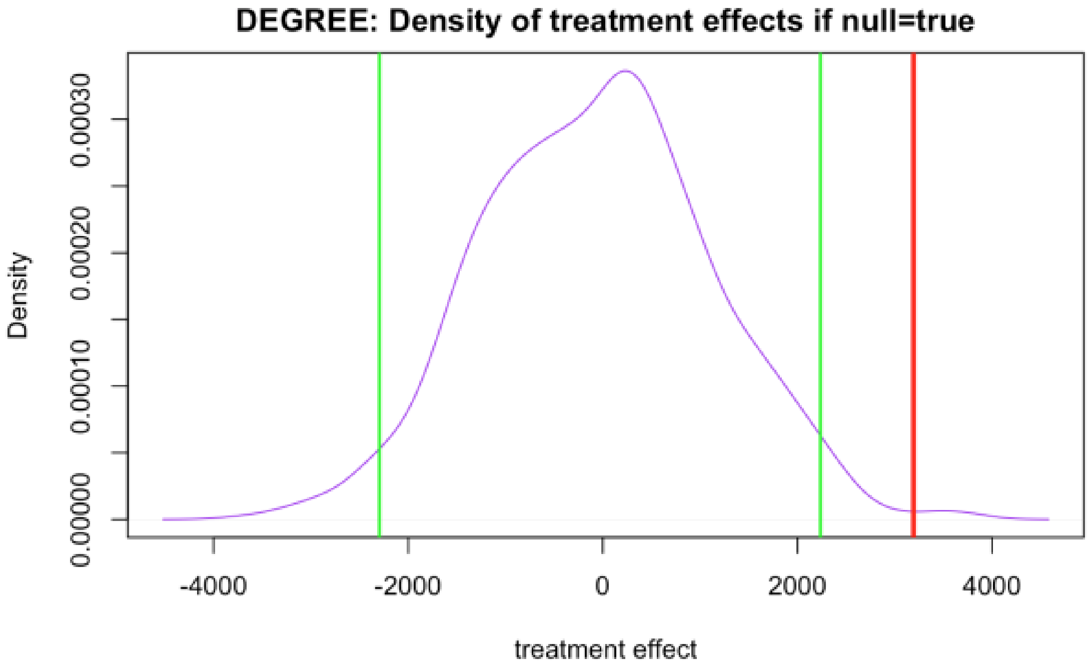


We can see that the maximum outcome if the null hypothesis is true from our simulation was $1586. Our treatment effect is 1794, which is larger than this maximum value! We can also see that visually in the plot. From this method too, we see that our observed treatment effect is indeed very unlikely to happen if our null hypothesis is true.

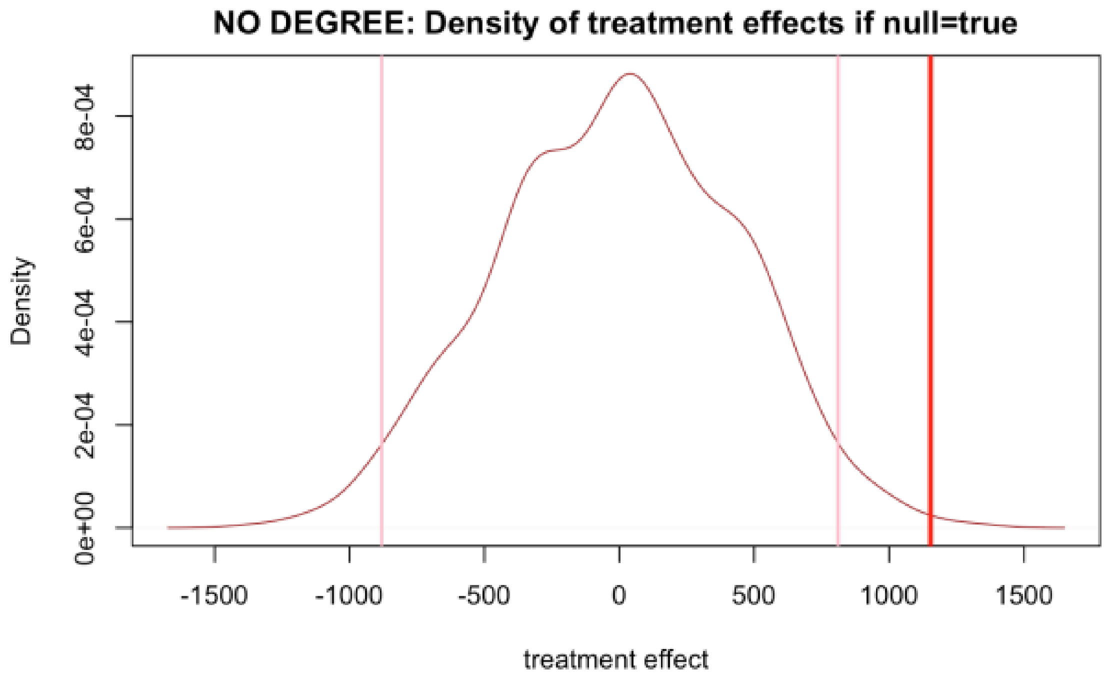
3.B – Conducting Fisher exact test separately for people with degree and without.

Here we applied the FET simulation of random assignments to groups and where our observed effect stand in relation to the distribution for those WITH DEGREE.

For people with a degree, the results show that the observed outcomes for people with degree are significantly higher than the hypothetical results if the null hypothesis was true. The difference between our true observed treatment effect ($3192) and the upper bound of the confidence interval of the null results (2248.331) is significant, at almost another 1000 dollars: 943.69.



The observed treatment effect for people with NO degree was: $1154.048.



From this we can see that the treatment effect for people without degree not only was much lower, but also closer to the upper limit of the confidence interval of the hypothetical null hypothesis results. Therefore, the treatment effect for people with no degree is both lower and less significant. For people with no degree, the difference between the observed outcomes (1154.048) to the upper bound of null hypothesis (889.27) results is 264.76. The proportion of this difference is 23% (264.76/1154.048). Whereas the difference for people with degree was more significant: 943.69; and the observed outcomes themselves were larger: 3192 average observed treatment effect. This is a proportion of about 30% (943.69/3192).

CONCLUSION

We analyzed results from an observational study taken at 1978, by Lalonde, through three different lenses: multiple (multivariate) linear regression, random forests, and Fisher's Exact Test. All three methods suggested that the program was indeed beneficial, and showed clearly that the program benefitted people with a high school degree far more than it benefitted people without a high school degree.

Conclusion: THE PROGRAM SHOULD BE TARGETED TOWARDS INDIVIDUALS WITH A HIGH SCHOOL DEGREE.

TECHNICAL APPENDIX

TECHNICAL DETAILS

Environment setup and data Prep

```
###### DATA PREP ######
#setting the workspace, libraries, data
#setwd("~/Google Drive/Academics (Tomer)/0-2017 Buenos Aires Sem2.2/CS112 DataScience/R_CS112") #(Yep. I know that's long)
library(Matching)

## Loading required package: MASS

## ## 
## ##  Matching (Version 4.9-2, Build Date: 2015-12-25)
## ##  See http://sekhon.berkeley.edu/matching for additional documentation.
## ##  Please cite software as:
## ##   Jasjeet S. Sekhon. 2011. ``Multivariate and Propensity Score Matching
## ##   Software with Automated Balance Optimization: The Matching package for R.''
## ##   Journal of Statistical Software, 42(7): 1-52. 
## ##

library(randomForest)

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

data("lalonde")
#head(lalonde)
#View(lalonde)
#summary(lalonde)
attach(lalonde)
```

1. MULTIVARIATE REGRESSION

Creating a Multivariate regression of real earnings

```
#multivariate regression of real earnings
lm2 <- lm(re78 ~ treat + educ + re75 + age + black + hisp + nodegr + u74 + u75 + married,
data=lalonde)
summary(lm2)

## 
## Call:
## lm(formula = re78 ~ treat + educ + re75 + age + black + hisp + 
##     nodegr + u74 + u75 + married, data = lalonde)
## 
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9113  -4456  -1609   3048  53306
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1285.0634  3448.6176   0.373   0.7096
## treat        1605.5218   640.1676   2.508   0.0125 *
## educ          371.7119   228.1403   1.629   0.1040
## re75            0.1202     0.1323   0.908   0.3643
## age            53.4331    45.8575   1.165   0.2446
## black       -2049.8534  1174.9980  -1.745   0.0818 .
## hisp          230.1306  1560.0865   0.148   0.8828
## nodegr       -200.5898   998.3066  -0.201   0.8408
## u74           622.3750  1059.5222   0.587   0.5572
## u75          -841.9766  1012.9377  -0.831   0.4063
## married      -155.1317   883.2236  -0.176   0.8607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6524 on 434 degrees of freedom
## Multiple R-squared:  0.05393,    Adjusted R-squared:  0.03213
## F-statistic: 2.474 on 10 and 434 DF,  p-value: 0.00689
```

```
coefficients(lm2)
```

```
##   (Intercept)         treat          educ          re75           age
##  1285.0634269  1605.5218160   371.7118909     0.1201764    53.4330546
##
```

CONFIDENCE INTERVALS for the variables of the multiple regression

```
confint(lm2,parm="treat",level = 0.95)
```

```
##          2.5 %   97.5 %
## treat 347.3076 2863.736
```

(1.d) Are estimated treatment effects higher for individuals with (or without) a high school degree?

Let's check. I will split the dataset by degree and nondegree

```
#splitting the dataset by degree and nondegree
data_degree <- subset(lalonde, nodegr == 0)
data_nodegree <- subset(lalonde, nodegr == 1)

#verifying and viewing the sub_dataframes
#head(data_nodegree)
#head(data_degree)
#dim(data_nodegree)
#dim(data_degree)
```

fitting regressions for degree and no degree

```
lm_nodgr <- lm(re78 ~ treat + educ + re75 + age + black + hisp + nodegr + u74 + u75 + mar
ried, data = data_nodegree)
lm_dgr <- lm(re78 ~ treat + educ + re75 + age + black + hisp + nodegr + u74 + u75 + marrie
d, data = data_degree)

#view regression results
summary(lm_nodgr)

##
## Call:
## lm(formula = re78 ~ treat + educ + re75 + age + black + hisp +
##     nodegr + u74 + u75 + married, data = data_nodegree)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7637  -4154  -1547   2902  52790
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.815e+03  3.089e+03   1.235   0.2177
## treat        1.052e+03  7.045e+02   1.493   0.1362
## educ         2.172e+02  2.286e+02   0.950   0.3426
## re75        -4.569e-02  1.373e-01  -0.333   0.7395
## age          5.103e+01  4.803e+01   1.062   0.2888
## black       -2.240e+03  1.465e+03  -1.529   0.1271
## hisp         2.955e+02  1.774e+03   0.167   0.8678
## nodegr              NA         NA      NA       NA
## u74          1.131e+03  1.134e+03   0.998   0.3190
## u75         -2.328e+03  1.101e+03  -2.114   0.0352 *
## married     -3.137e+00  9.762e+02  -0.003   0.9974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6250 on 338 degrees of freedom
## Multiple R-squared:  0.04494,    Adjusted R-squared:  0.01951
## F-statistic: 1.767 on 9 and 338 DF,  p-value: 0.07337

summary(lm_dgr)

##
## Call:
## lm(formula = re78 ~ treat + educ + re75 + age + black + hisp +
##     nodegr + u74 + u75 + married, data = data_degree)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11528  -4292  -1771   3005  25551
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.851e+04  1.249e+04  -1.482  0.14199
## treat        2.283e+03  1.556e+03   1.467  0.14596
## educ         1.779e+03  1.037e+03   1.716  0.08980 .
## re75         1.142e+00  4.167e-01   2.741  0.00744 **
## age          3.886e+01  1.330e+02   0.292  0.77079
## black       -2.405e+03  2.071e+03  -1.161  0.24875
## hisp        -4.106e+03  4.277e+03  -0.960  0.33976
## nodegr              NA         NA      NA       NA
## u74         -5.559e+02  2.828e+03  -0.197  0.84464
## u75          3.557e+03  2.525e+03   1.409  0.16246
## married     -7.485e+02  2.031e+03  -0.368  0.71344
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7217 on 87 degrees of freedom
## Multiple R-squared:  0.1735, Adjusted R-squared:  0.08805
## F-statistic:  2.03 on 9 and 87 DF,  p-value: 0.04519
```

```
#view specifically the coefficients
coefficients(lm_nodgr)
```

```
##   (Intercept)         treat          educ          re75           age
##  3.814542e+03  1.052133e+03  2.172090e+02 -4.569302e-02  5.102826e+01
##         black          hisp        nodegr           u74           u75
## -2.239926e+03  2.954593e+02            NA  1.131487e+03 -2.328187e+03
##       married
## -3.137263e+00
```

```
coefficients(lm_dgr)
```

```
##   (Intercept)         treat          educ          re75           age
## -18508.014283   2283.077018   1779.077914      1.142075     38.857032
##         black          hisp        nodegr           u74           u75
##  -2404.676439  -4105.514807            NA   -555.865097   3556.961281
##       married
##   -748.466648
```

```
#view confidence intervals of coefficients
confint(lm_nodgr)
```

```
##                   2.5 %       97.5 %
## (Intercept) -2260.5973051 9889.6803384
## treat        -333.5791093 2437.8453263
## educ         -232.3758727  666.7938116
## re75           -0.3157609    0.2243748
## age           -43.4477085  145.5042237
## black       -5120.8272776  640.9752347
## hisp        -3194.3042899 3785.2229476
## nodegr                 NA           NA
## u74         -1098.5292995 3361.5025872
## u75        -4494.3525323 -162.0216133
## married    -1923.3372284 1917.0627017
```

```
confint(lm_dgr)
```

```
##                  2.5 %      97.5 %
## (Intercept) -4.333247e+04 6316.438869
## treat       -8.100472e+02 5376.201243
## educ        -2.821153e+02 3840.271126
## re75         3.137677e-01    1.970381
## age         -2.254143e+02  303.128351
## black       -6.520850e+03 1711.496942
## hisp        -1.260643e+04 4395.397110
## nodegr                 NA          NA
## u74         -6.177262e+03 5065.531811
## u75         -1.461399e+03 8575.321405
## married     -4.786137e+03 3289.203685
```

Confidence intervals for people with No Degree:

treat   -333.58   2437.84

Confidence intervals for people with Degree:

treat   -810.05   5376.20

Therefore, we see that there is a big overlap between -810.05 and 2437.84. So while that the treatment effect is different on average between groups, the true mean effect might actually be the same, or even smaller for people with a degree than people without a degree. Additionally, the confidence interval for people with degree is much wider, indicating that we have a less certainty about the treatment effect for people with degree, which decreases our certainty of results.

*See detailed analysis in text body.

2. RANDOM FOREST REGRESSION

2.a - fit random forest model, fitting 78 onto the same predictors as before

```
forest <- randomForest(re78 ~ treat + educ + re75 + age + black + hisp + nodegr + u74 + u
75 + married, data = lalonde, importance=TRUE, ntree=500)
summary(forest)
```

```
##                 Length Class  Mode
## call              5    -none- call
## type              1    -none- character
## predicted       445    -none- numeric
## mse             500    -none- numeric
## rsq             500    -none- numeric
## oob.times       445    -none- numeric
## importance       20    -none- numeric
## importanceSD     10    -none- numeric
## localImportance   0    -none- NULL
## proximity         0    -none- NULL
## ntree             1    -none- numeric
## mtry              1    -none- numeric
## forest           11    -none- list
## coefs             0    -none- NULL
## y               445    -none- numeric
## test              0    -none- NULL
## inbag             0    -none- NULL
## terms             3    terms  call
```

```
print(forest)
```

```
##
## Call:
##  randomForest(formula = re78 ~ treat + educ + re75 + age + black +      hisp + nodegr +
u74 + u75 + married, data = lalonde, importance = TRUE,      ntree = 500)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 46348612
##                     % Var explained: -5.63
```

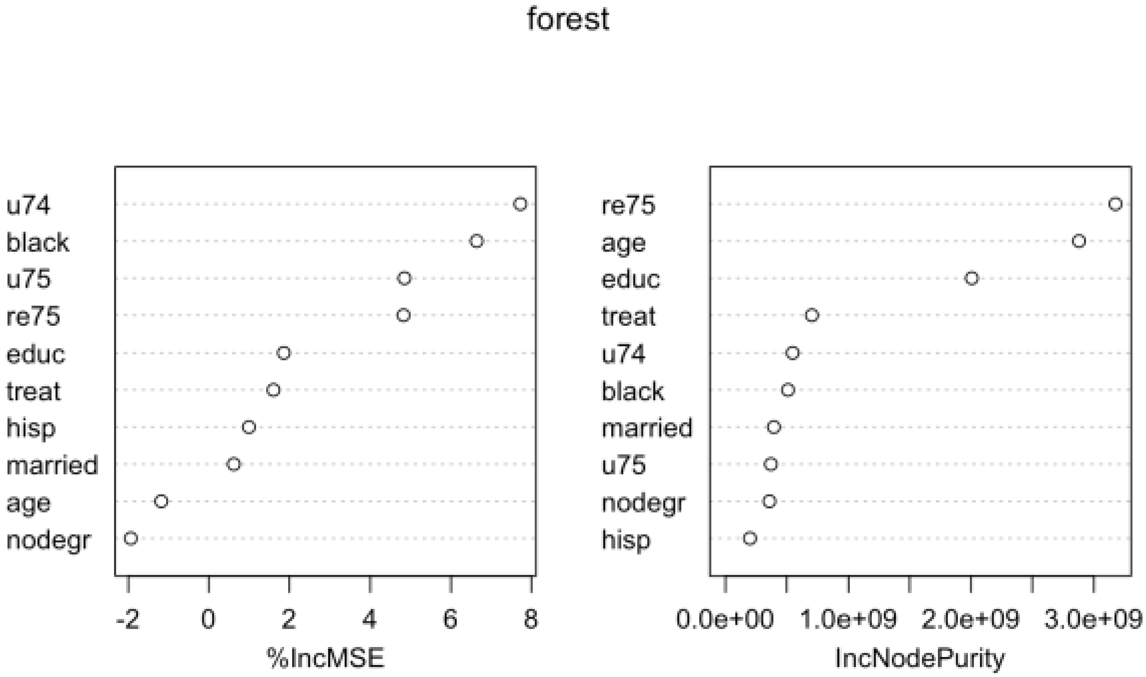
```
#plot(forest)
```

2.a - predictor importance table

We see by the variable importance table and plot how "important" was each variable. Meaning, if we took out this variable, by how much percent would are Mean Squared Error increase? We see that unemployment in 74 ,black, unemployment in 75 and real earnings 75 were more instrumental for our Forest's accuracy. interestingly, the treatment variable is only after these and years of education 6th in importance, and has relatively low inc-node purity.

```
importance(forest)
```

```
##           %IncMSE IncNodePurity
## treat   2.4149342     675228821
## educ    1.4712992    1880170955
## re75    5.1054244    3022319952
## age     0.4085220    2811240656
## black   9.3013140     485228268
## hisp    2.2320814     222934469
## nodegr -0.2241512     336945600
## u74     7.6235317     549819977
## u75      4.2463321    347919709
## married  3.3076306    406900065

varImpPlot(forest)
```



2.b - predict treatment effect for the treatment units

```
#create dataset of only treated units, without re78
data_treated <- lalonde[treat == 1,]
data_treated_cf <- lalonde[treat == 1,]

  #dim(data_treated)
  #head(data_treated_cf)
```

2.b - We'll predict counterfactual estimated for what would be the real earnings for the treated units if they weren't treated? Then we'll calculate the treatment effect based on the test statistic of the mean difference between each unit's potential outcomes, if treated (as in reality) or if it weren't treated (estimated).

```
#change all "treated = 1" to 0 (control)
data_treated_cf$treat <- 0
#head(data_treated_cf)

#use predict function basing on the forest model fitted upon the entire dataset


rf_predicted_cf <- predict(forest, data_treated_cf)

#estimated treatment effect for only the treated units, based on potential counterfactual outc
omes for each unit.
```

```
#("What would be the difference in earnings 1978 for the treated units between reality where
they've been treated, to if they weren't treated?")
mean(data_treated_cf$re78 - rf_predicted_cf)

## [1] 1334.931
```

We see the treatment effect for the all treated units is $1273.414. Meaning, the treated units are estimated to have gained $1273.414 on average by the program.

2.C - Degree / vs NoDegree Random Forest

Now we want to test the same: the difference in potential outcomes for the treated people, but separately for people who have a degree and for people without a degree.

```
# create dataset of only treated units, without re78
treated_yesdegree <- subset(data_treated, nodegr == 0)
treated_nodegree <- subset(data_treated, nodegr == 1)

#view and verify our subdatasets look right, if we want:
#head(data_degree)
#head(data_nodegree)
#dim(data_degree)
#dim(data_nodegree)
```

2.C - Predict counterfactuals for each subgroup

First, we will create new datasets that will be the same dataset of the treated units, but would "fake" the same units to be in the control group, by changing "treat" from "1" to "0".

```
#create new datasets that will contain counterfactuals
treated_yesdg_cf <- subset(treated_yesdegree, nodegr == 0)
treated_nodg_cf <- subset(data_treated, nodegr == 1)

#change all "treated = 1" to 0 (control)
treated_yesdg_cf$treat <- 0
treated_nodg_cf$treat <- 0
```

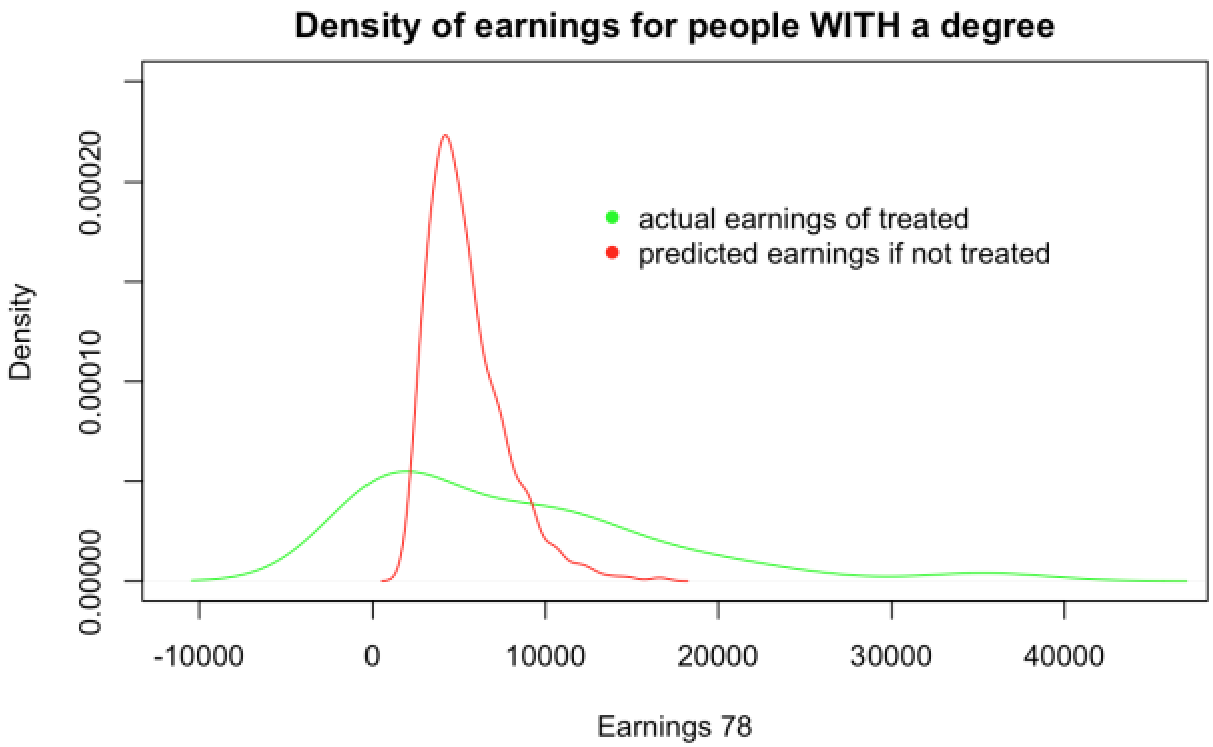
Now we will predict the earnings in 1978 for each subgroup.

```
#use predict function, basing on the forest model fitted upon the entire dataset, what would b
e the results of
predicted_re78_yesdg_treated_cf <- predict(forest,data=treated_yesdg_cf)
predicted_re78_nodg_treated_cf <- predict(forest, data=treated_nodg_cf)
```

2.C1 - WITH DEGREE COMPARISON.

Comparing the potential outcomes for the people WITH DEGREE who were treated (in the program). This means we estimate how much would they have gained from wether they haven't been given the treatment.

```
#treatment effect of treated people with degree

#viewing each mean
print("average earnings in 1978 for treated with degree")

## [1] "average earnings in 1978 for treated with degree"

mean(treated_yesdegree$re78)

## [1] 8046.521

print("predicted mean earnings in 1978 for treated people with degree IF THEY HADN'T B
EEN TREATED")

## [1] "predicted mean earnings in 1978 for treated people with degree IF THEY HADN'T B
EEN TREATED"

mean(predicted_re78_yesdg_treated_cf)

## [1] 5383.241

#treatment effect (only treated)
print("treatment effect for treated people with degree")

## [1] "treatment effect for treated people with degree"

mean(treated_yesdegree$re78) - mean(predicted_re78_yesdg_treated_cf)

## [1] 2663.28

#plot WITH DEGREE counterfactual comparison
plot(density(treated_yesdegree$re78), xlim = c(-11000,46000),ylim = c(0,0.00025), col="
green", main = "Density of earnings for people WITH a degree", ylab = "Density", xlab = "Ear
nings 78")
lines(density(predicted_re78_yesdg_treated_cf),col="red")
legend(x=42000, y=0.00025, xjust=1, col = c("green","red"), legend = c("actual earnings of
treated", "predicted earnings if not treated"),  bty="n", pch= c(16, 16) )
```

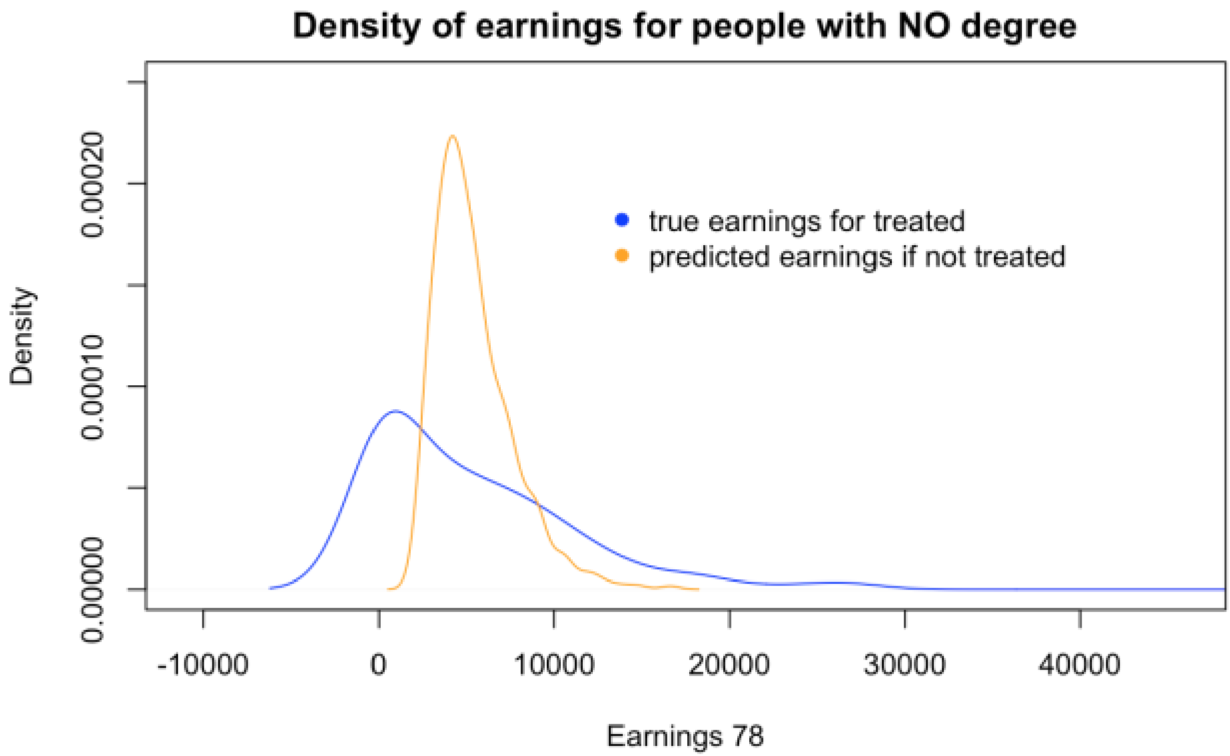


So, the treatment effect for treated people WITH DEGREE is: $2647.82

2.C2 - WITH NO DEGREE COMPARISON

```
#treatment effect of treated people without degree
print("average earnings in 1978 for treated with NO degree")

## [1] "average earnings in 1978 for treated with NO degree"

mean(treated_nodegree$re78)

## [1] 5649.464

print("predicted mean earnings in 1978 for treated people with NO degree IF THEY HADN'
T BEEN TREATED")

## [1] "predicted mean earnings in 1978 for treated people with NO degree IF THEY HADN
'T BEEN TREATED"

mean(predicted_re78_nodg_treated_cf)

## [1] 5383.241

print("treatment effect for treated people with NO degree")

## [1] "treatment effect for treated people with NO degree"

mean(treated_nodegree$re78) - mean(predicted_re78_nodg_treated_cf)

## [1] 266.2234
```

```
#plot WITHOUT DEGREE counterfactual comparison
plot(density(treated_nodegree$re78), xlim = c(-11000,46000),ylim = c(0,0.00025), col="blue", main = "Density of earnings for people with NO degree", ylab = "Density", xlab = "Earnings 78")
lines(density(predicted_re78_nodg_treated_cf),col="orange")
legend(x=40000, y=0.00020, xjust=1, col = c("blue","orange"), legend = c("true earnings for treated", "predicted earnings if not treated"),  bty="n", pch= c(16, 16) )
```



So, the treatment effect for treated people WITH NO DEGREE is: $250.76 This is MUCH lower, almost miniscule, compared to the treatment effect for treated people WITH DEGREE, which was: $2647.82. Meaning, people with degree gained from the treatment, on average, **$2397.06** more than people without degree gained from the treatment. ($2647.82 vs $250.76).

3. IMPLEMENT A FISHER EXACT TEST (FET) USING THE SHARP NULL HYPOTHESIS OF NO TREATMENT EFFECT FOR ANY UNIT.

1. Defining a test statistic for estimating the treatment effect and evaluating the null hypothesis: ** We define the test statistic as the mean of earnings for all units minus mean of all units untreated (y(1) ??? y(0)) **

Calculating the actual observed treatment effect between the original groups:

```
original_observed_meandiff <- (mean(lalonde[treat == 1,]$re78) - mean(lalonde[treat == 0
,]$re78))
original_observed_meandiff

## [1] 1794.343
```

The original observed treatment effect between the original groups is: $1794.343.

2. Defining a sharp null hypothesis (hypothesis regarding the size of the treatment effect on each unit): ** We will define the null hypothesis as "treatment effect = 0"" of absolutely no effect of treatment: (Yi(0) = Yi(1) for all units). **

To implement Fisher's exact test, we will take random assignments to treatment, calculate the observed average treatment effect for each random assignment, and store these results in a storage list.

Now we will make many random assignments of the units to control and treatment groups.

For each random assignment: a. Calculate treatment effect: Average all Y(1)s -minus all Y(0) results. (this would depend on the initial values; if I happen to assign to the treatment all units with smaller values, and the larger values to the control, I would get a negative average treatment effect even if for each unit individually is increasing with treatment.)

```
 b. Save this treatment effect to a list of treatment effects (1 for each assignment).
```

Here we create the assignment function, which randomizes into treatment and control groups, and returns the observed average treatment effect for that assignment.

```
# Assignment function: randomizes groups and returns the observed average treatment effect

rand_assign_mean <- function() {
  #randomizing units to be treated: half of the people
  randtreat <- sample(nrow(lalonde),nrow(lalonde)/2,replace = FALSE)

  #taking the rest to be control
  randcontrol <- nrow(lalonde)-randtreat

  #assigning vectors of outcomes
  randcontrol_outcomes <- lalonde[randcontrol,]$re78
  randtreatmnet_outcomes <- lalonde[randtreat,]$re78

  #calculate the observed mean treatment effect
  return(mean(randtreatmnet_outcomes) - mean(randcontrol_outcomes))
}
```

Now we will iterate through many of these random assignments:

```
### *this code snippet is assisted by Ben Imadali's code.

#initiating a storage list vector (empty for now, later will store our results)
storage <- NULL

# Iterating the Assignment function
iter.RI <- function(iterations = 500) {
  for (i in 1:iterations)
  {storage[i] <- rand_assign_mean()
  }
  return(storage)
}

results <- iter.RI()
head(results)

## [1]  353.75971  261.97961   18.65697 -263.78239  223.03249  -81.73897
```

Exploring the results

Now we will observe the distribution of our resulting treatment effects. Since each random assignment to groups would yield a different treatment effect, even if the treatment has no actual effect (true treatment effect is = 0), we would have a range of these observed effects given the treatment effect is 0. We want to understand where does our REAL observed treatment effect (from the original experiment) stands in relation to these results, then we could understand HOW LIKELY was it to observe our treatment effect, given the null hypothesis is true.

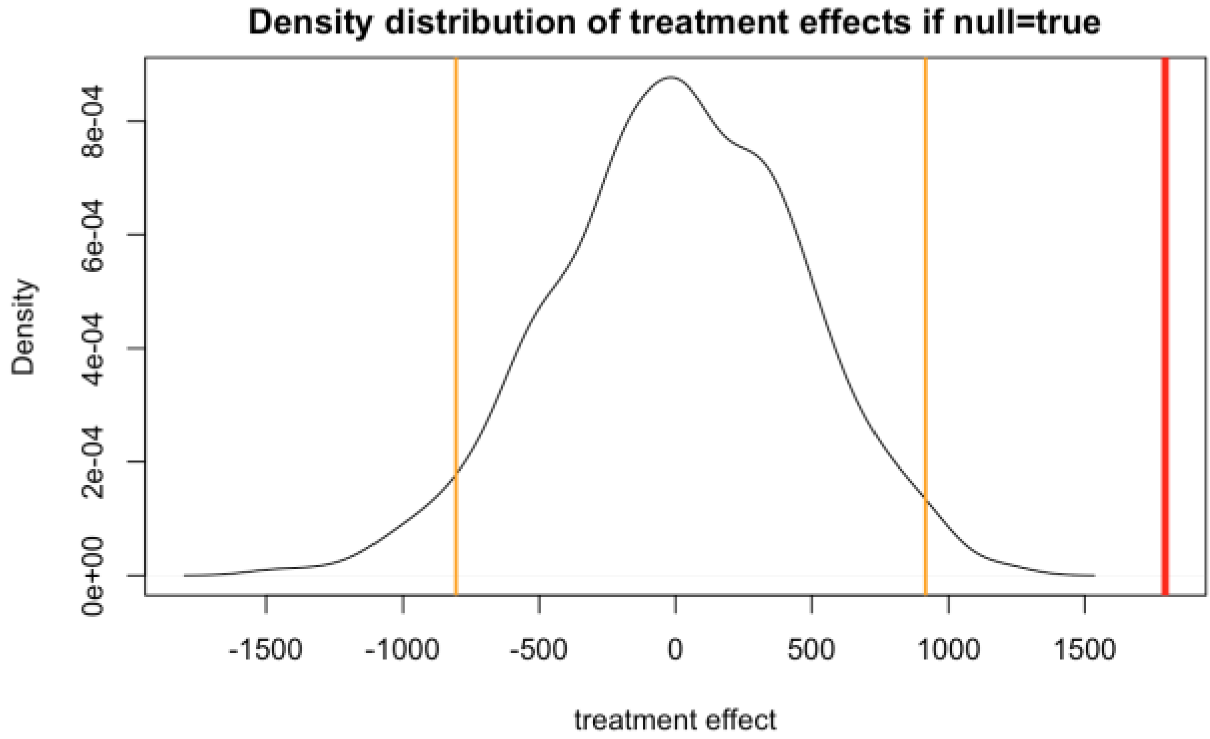
first, let's see our confidence intervals for the results:

```
quantile(results, prob = c(0.025, 0.975))

##      2.5%     97.5%
## -851.0583  771.0822
```

we see our confidence interval of 95% (between the 2.5% and 97.5%) is: {-805.3602 to 915.4775 }.

Let's explore the data a bit more, and plot the distribution of potential treatment effects if the null hypothesis was true. The actual observed treatment effect between the original groups was, as we saw before: 1794.343

```
#plot the distribution of potential treatment effects if the null hypothesis was true
plot(density(results), xlim = c(-1800,1800), main = "Density distribution of treatment effe
cts if null=true", ylab = "Density", xlab = "treatment effect")



#Where does our original treatment effect stand upon on that distribution? the red line:
abline(v = 1794.343, lwd = 4, col = "red")

max(results)

## [1] 1671.181

#our confidence interval plotted upon this distriubtion
abline(v = -805.3602, lwd = 2, col = "orange")
abline(v =  915.4775, lwd = 2, col = "orange")
```



We can actually see that the maximum outcome if the null hypothesis is true from our simulation was $1586. Our treatment effect is 1794, which is larger than this maximum value! We can also see that visually in the plot. From this method too, we see that our observed treatment effect is indeed very unlikely to happen if our null hypothesis is true.

3.B - Fisher exact test separately for people with degree and without.

Observed mean ### 3.B.1 - FET for people WITH DEGREE

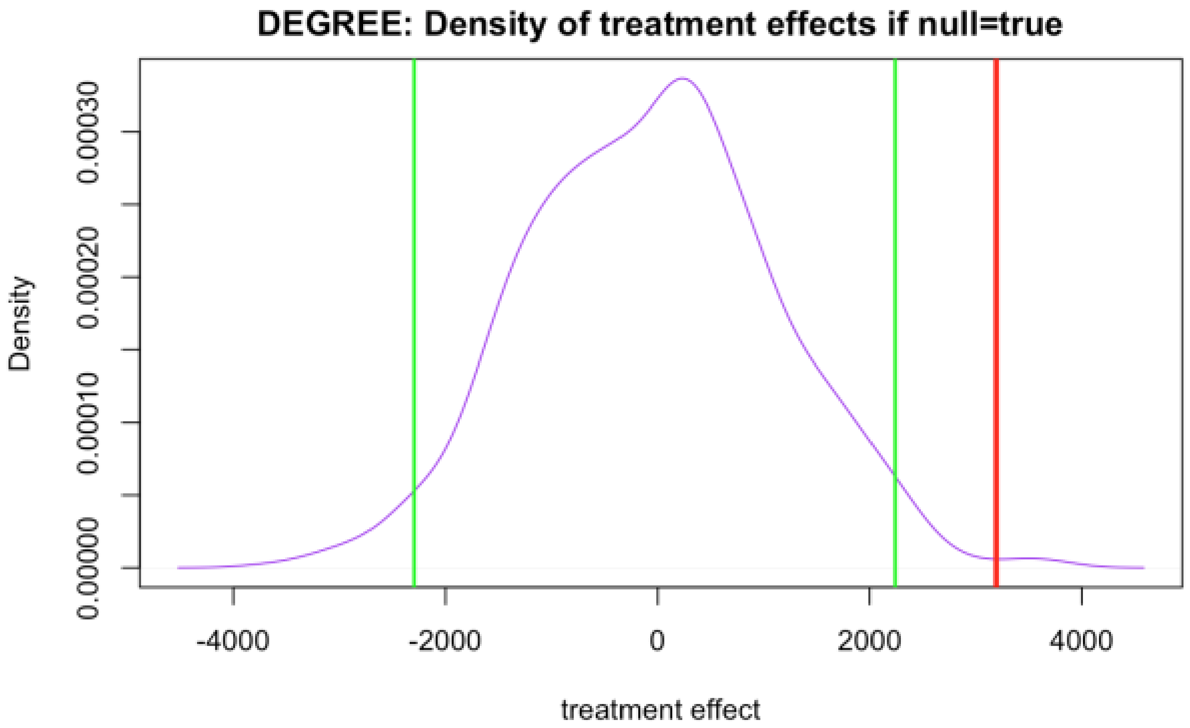
```
#data_degree

#treatment
degree_treatment_re <- subset(data_degree, treat == 1, select=re78)
#removing NA values
degree_treatment_re <- degree_treatment_re[!is.na(degree_treatment_re)]
#control
degree_control_re <- subset(data_degree, treat == 0, select=re78)
#removing NA values
degree_control_re <- degree_control_re[!is.na(degree_control_re)]

degree_observed_meandiff <- mean(degree_treatment_re) - mean(degree_control_re)
degree_observed_meandiff

## [1] 3192.026
```

Observed treatment effect for people with degree was $3192.

Now we'll apply the FET simulation of random assignments to groups and where our observed effect stand in relation to the distribution for those WITH DEGREE.

```
# Assignment function: randomizes groups and returns the observed average treatment effect

#data_degree


rand_assign_mean_degree <- function() {
  #randomizing units to be treated: half of the people
  randtreat <- sample(nrow(data_degree),nrow(data_degree)/2,replace = FALSE)

  #taking the rest to be control
  randcontrol <- nrow(data_degree)-randtreat

  #assigning vectors of outcomes
  randcontrol_outcomes <- data_degree[randcontrol,]$re78
  randcontrol_outcomes <- randcontrol_outcomes[!is.na(randcontrol_outcomes)]
  randtreatmnet_outcomes <- data_degree[randtreat,]$re78
  randtreatmnet_outcomes <- randtreatmnet_outcomes[!is.na(randtreatmnet_outcomes)]

  #calculate the observed mean treatment effect
  return ( mean(randtreatmnet_outcomes) - mean(randcontrol_outcomes) )
}
```

```
#initiating a storage list vector (empty for now, later will store our results)
storage_degree <- NULL

rand_assign_mean_degree()

## [1] -1427.096

# Iterating the Assignment function
iter.RI <- function(iterations = 1000) {
  for (i in 1:iterations)
  {storage_degree[i] <- rand_assign_mean_degree()
  }
  return(storage_degree)
}

results_degree <- iter.RI()
head(results_degree)

## [1] -2192.02079 -1702.06015 -1476.21681    33.96369  -396.87708  2075.32681

#confidence interval for results
quantile(results_degree, prob = c(0.025, 0.975))

##      2.5%     97.5%
## -2267.001  2499.471

# confidence interval: -2349.641  2453.530

#plot the distribution of potential treatment effects if the null hypothesis was true
plot(density(results_degree), col="purple", main = "DEGREE: Density distribution of treat
ment effects for people with DEGREE if null=true", ylab = "Density", xlab = "treatment effect
")

#Where does our original treatment effect stand upon on that distribution? the red line:
abline(v = degree_observed_meandiff, lwd = 3, col = "red")

max(results)

## [1] 1671.181

#our confidence interval plotted upon this distriubtion
abline(v = quantile(results_degree, prob = c(0.025, 0.975)), lwd = 2, col = "green")
```



```
#degree_observed_meandiff
#quantile(results_degree, prob = c(0.025, 0.975))[2]
deg_diff_from_conf <- degree_observed_meandiff - quantile(results_degree, prob = c(0.025,
0.975))[2]
deg_diff_from_conf

##    97.5%
## 692.5557
```

For people with a degree, the results show that the observed outcomes for people with degree are significantly higher than the hypothetical results if the null hypothesis was true. The difference between our true observed treatment effect ($3192) and the upper bound of the confidence interval of the null results (2248.331) is significant, at almost another 1000 dollars: 943.69.

3.B.3 - FET for people with NO DEGREE

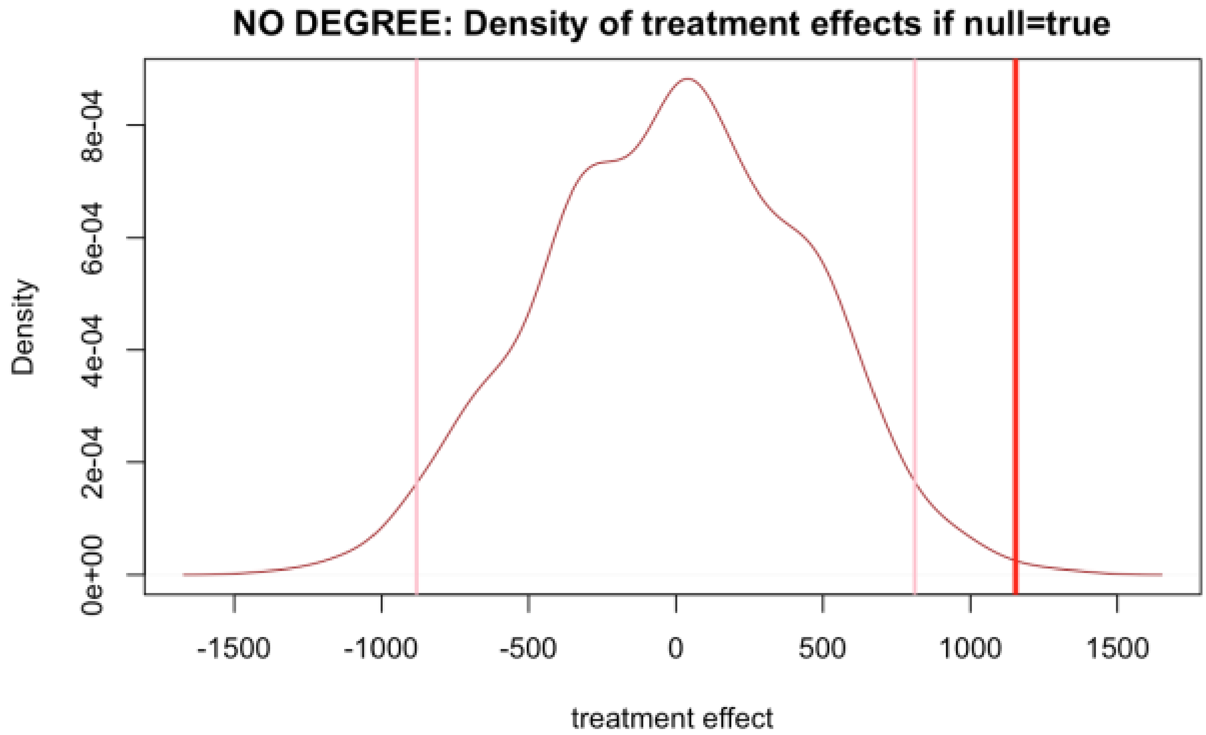
```
#data_nodegree

#treatment
nodeg_treatment_re <- subset(data_nodegree, treat == 1, select=re78)
#removing NA values
nodeg_treatment_re <- nodeg_treatment_re[!is.na(nodeg_treatment_re)]
#control
nodeg_control_re <- subset(data_nodegree, treat == 0, select=re78)
#removing NA values
nodeg_control_re <- nodeg_control_re[!is.na(nodeg_control_re)]
```

```
nodeg_observed_meandiff <- mean(nodeg_treatment_re) - mean(nodeg_control_re)
nodeg_observed_meandiff

## [1] 1154.048
```

The observed treatment effect for people with NO degree was: $1154.048.

```
# Assignment function: randomizes groups and returns the observed average treatment effect


rand_assign_mean_nodeg <- function() {
  #randomizing units to be treated: half of the people
  randtreat <- sample(nrow(data_nodegree),nrow(data_nodegree)/2,replace = FALSE)

  #taking the rest to be control
  randcontrol <- nrow(data_nodegree)-randtreat

  #assigning vectors of outcomes
  randcontrol_outcomes <- data_nodegree[randcontrol,]$re78
  randcontrol_outcomes <- randcontrol_outcomes[!is.na(randcontrol_outcomes)]
  randtreatmnet_outcomes <- data_nodegree[randtreat,]$re78
  randtreatmnet_outcomes <- randtreatmnet_outcomes[!is.na(randtreatmnet_outcomes)]

  #calculate the observed mean treatment effect
  return ( mean(randtreatmnet_outcomes) - mean(randcontrol_outcomes) )
}



#initiating a storage list vector (empty for now, later will store our results)
storage_nodeg <- NULL

rand_assign_mean_nodeg()

## [1] 485.553

# Iterating the Assignment function
iter.RI <- function(iterations = 1000) {
  for (i in 1:iterations)
  {storage_nodeg[i] <- rand_assign_mean_nodeg()
  }
  return(storage_nodeg)
}

results_nodeg <- iter.RI()
head(results_nodeg)

## [1] -228.9128 -934.8457 -295.8479  189.6410  248.2852 -163.2239

#confidence interval for results
quantile(results_nodeg, prob = c(0.025, 0.975))
```

```
##      2.5%     97.5% 
## -869.7289  933.4408

# confidence interval: -2349.641  2453.530

#plot the distribution of potential treatment effects if the null hypothesis was true
plot(density(results_nodeg), col="brown", main = "NO DEGREE: Density distribution of tre
atment effects for people with NO DEGREE if null=true", ylab = "Density", xlab = "treatment
effect")

#Where does our original treatment effect stand upon on that distribution? the red line:
abline(v = nodeg_observed_meandiff, lwd = 3, col = "red")

#max(results)

#our confidence interval plotted upon this distriubtion
abline(v = quantile(results_nodeg, prob = c(0.025, 0.975)), lwd = 2, col = "pink")
```



```
nodeg_observed_meandiff

## [1] 1154.048

quantile(results_nodeg, prob = c(0.025, 0.975))[2]

##    97.5% 
## 933.4408

nodeg_observed_meandiff - quantile(results_nodeg, prob = c(0.025, 0.975))[2]
```

```
##    97.5%
## 220.6069
```

From this we can see that the treatment effect for people without degree not only was much lower, but also closer to the upper limit of the confidence interval of the hypothetical null hypothesis results. Therefore, the treatment effect for people with no degree is both lower and less significant. For people with no degree, the difference between the observed outcomes (1154.048) to the upper bound of null hypothesis (889.27) results is 264.76. The proportion of this difference is 23% (264.76/1154.048). Whereas the difference for people with degree was more significant: 943.69; and the observed outcomes themselves were larger: 3192 average observed treatment effect. This is a proportion of about 30% (943.69/3192).

```
943.69/3192

## [1] 0.2956422

264.76/1154.048

## [1] 0.2294185
```

CONCLUSION

All our methods suggest that the program is significant. Assuming the costs for running it are reasonable for making an increase of ~$1500 - we should run the program. The effects were MUCH more useful for people WITH degree than for people WITHOUT degree. However, the effects for people without degree were still positive, even if small. Therefore, I would recommend definitely running the program for participants with degree. If the program still has funds left after covering all people with a degree, they could start accepting people with no degree. However, they should probably try to adapt the program to this segment, since it apparently is not the most effective program for them. Alternatively, if they have a better working program with a more significant positive impact than this impact on people without a degree, they can simply take these extra funds and reallocate them for that other program.