Effects of Job Training on Wages

Ana Belen Barcenas J. | Prajwal Vijendra

10/20/2018

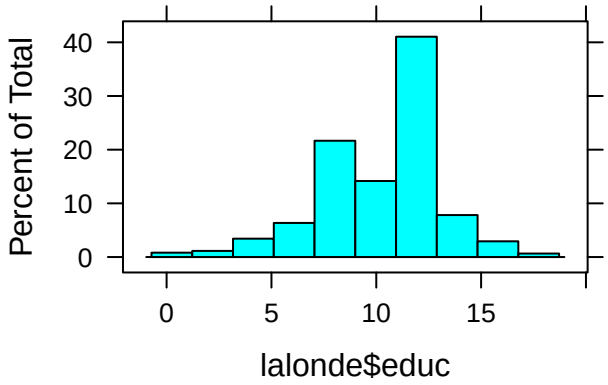
Let's analyze the data and transform or create variables that will be useful for the models

The race variable is divided in two categories. We will create a categorical variable = 1 if the person is black, = 2 if hispanic and = 3 otherwise:

```
lalonde$race <- ifelse(lalonde$black == 1, 1, 3)
lalonde$race <- ifelse(lalonde$hisp == 1, 2, lalonde$race)
```

Let's look at distribution of people for each edcuation level and also caluclate what is the mean of the wage of each education level.

```
histogram(lalonde$educ)
```




```
table(lalonde$educ)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##   3   2   2   5   9  12  12  27  62  71  87  95 157  27  21  10   8   2
##  18
##   2
```

```
tapply( lalonde$re78,lalonde$educ, mean)
```

```
##         0         1         2         3         4         5         6
##  3381.038  9732.305  5348.631  6445.253  2945.697  7946.163  4567.775
##         7         8         9        10        11        12        13
##  3637.468  5293.766  5901.474  6660.376  6733.042  7610.109  8076.161
##        14        15        16        17        18
## 11244.395  6598.415 11589.149 15584.105  9113.880
```

Based on the histogram and distribution, it would be better to bin the education into four category: middle school, high school, undergraduation, postgraduation.

Till Middle School: 0; Middle School to High School: 1; High School to Undergraduation:2; Undergraduation to Postgraduation: 3

```
lalonde$educ_cat <- ifelse(lalonde$educ %in% c(0,1,2,3,4,5,6,7), 0,
                           ifelse(lalonde$educ %in% c(8,9,10,11,12), 1, ifelse(lalonde$educ %in% c(
```

Let's see if this new variable is correlated with nodegree as well as the numerical educ var:

```
cor(lalonde$nodegree, lalonde$educ_cat)
```

```
## [1] -0.4908882
```

```
cor(lalonde$nodegree, lalonde$educ)
```

```
## [1] -0.7014519
```

The correlation is not that high, nothing to worry about.

We do not have experience information. We can create a proxy called "Potential experience". However, we should take into account that the potential experience could be highly correlated with education since it is calculated as PotencialExperience = Age - Education(years) - 6. Also, we should compare how well it explain the wage compared with "Age" since we can not use both vars.

```
lalonde$potential_exp <- lalonde$age - lalonde$educ - 6
lalonde$potential_expc = lalonde$potential_exp - mean(lalonde$potential_exp)
```

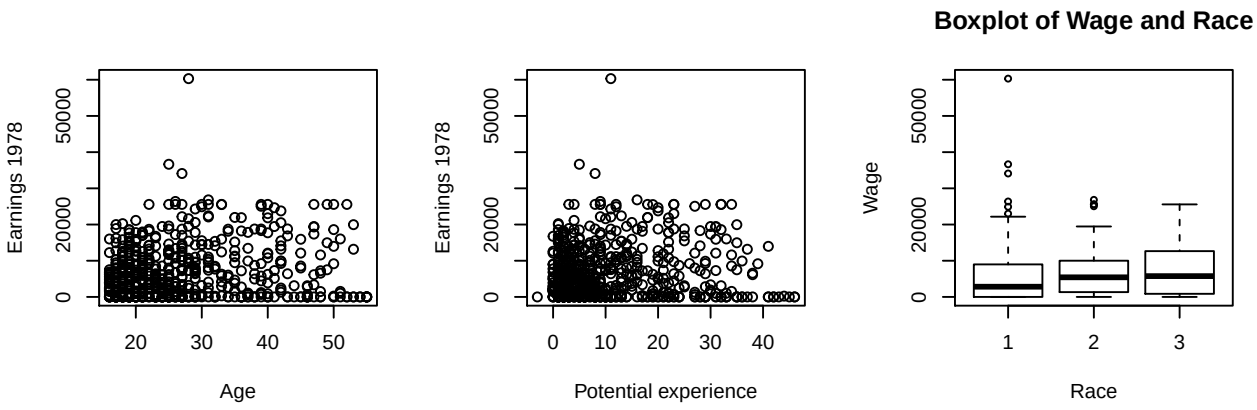
Now, let's mean center the explanatory varibales that can not be zero

```
lalonde$age_cent = lalonde$age - mean(lalonde$age)
lalonde$re74_cent = lalonde$re74 - mean(lalonde$re74)
lalonde$re75_cent = lalonde$re75 - mean(lalonde$re75)
```

1) IMPACT OF TREATMENT ON WAGES:

EXPLORATORY DATA ANALYSIS

```
par(mfrow=c(1,3))
plot(lalonde$age, lalonde$re78, xlab = "Age", ylab = "Earnings 1978")
plot(lalonde$potential_exp, lalonde$re78, xlab = "Potential experience", ylab = "Earnings 1978")
boxplot(re78~race, data = lalonde, xlab = "Race", ylab = "Wage", main = "Boxplot of Wage and Race")
```




```
par(mfrow=c(1,3))
boxplot(re78~treat, data = lalonde, xlab = "Treat", ylab = "Wage", main = "Boxplot of Wage and Treat")
boxplot(re78~married, data = lalonde, xlab = "Married", ylab = "Wage", main = "Boxplot of Wage and Marri
boxplot(re78~educ_cat, data = lalonde, xlab = "Education_Grouped", ylab = "Wage", main = "Boxplot of Wag
```

Looking at the scatterand box plot, it does not show strong linearity and constant variance. On the other hand, the literature says that the experi ence has an non linear effect on wages. Thus, we will use potential experience and potential experience squared in our analysis.

```
lalonde$potential_expc_sq <-  lalonde$potential_expc^2
```

We checked interactions by plotting lattice plots of race, wage, experience, treat, and education. Some interactions makes sence according to the science behind the problem. However, we did not find any signs of any interactions between the variables.

LET'S BUILD THE LINEAR MODEL

Once we have done the exploratory analysis and there is no evidence of interactions, we will build simple model. Moreover, given the high correlation between education and nodegree variable, we will not include the nodegree variable in the analysis.

```
wages.lm <- lm(re78~ as.factor(race) + as.factor(married) + as.factor(educ_cat) + re74_cent +re75_cent
summary(wages.lm)
```

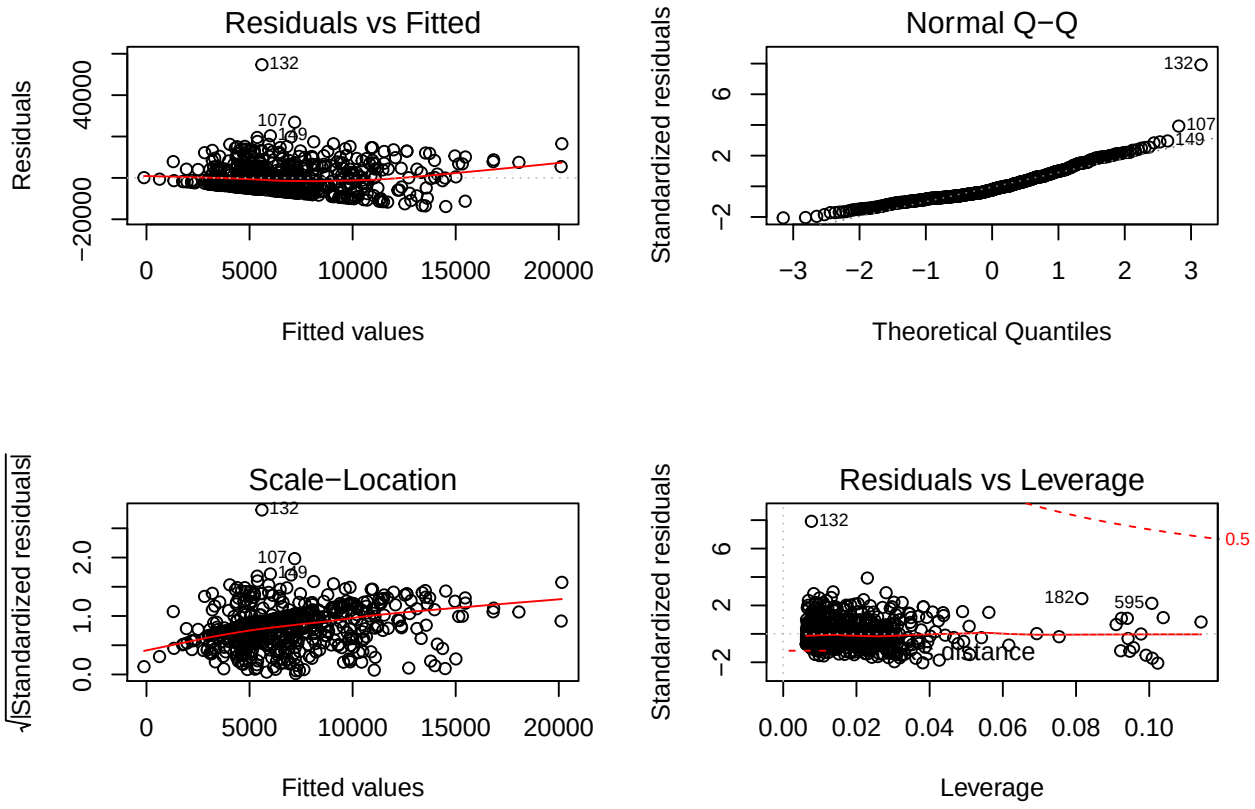
```
##
## Call:
## lm(formula = re78 ~ as.factor(race) + as.factor(married) + as.factor(educ_cat) +
##     re74_cent + re75_cent + as.factor(treat) + potential_expc +
##     potential_expc_sq, data = lalonde)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13846  -4791  -1585   4046  54708
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4331.09584 1219.81680   3.551 0.000414 ***
## as.factor(race)2     1536.42746 1014.72458   1.514 0.130517
## as.factor(race)3     1293.70060  767.39855   1.686 0.092347 .
## as.factor(married)1   210.69071  720.03608   0.293 0.769920
## as.factor(educ_cat)1 1454.22515  988.37954   1.471 0.141726
## as.factor(educ_cat)2 3499.23829 1329.85688   2.631 0.008724 **
## as.factor(educ_cat)3 5712.42461 2253.74960   2.535 0.011509 *
## re74_cent               0.30079    0.05764   5.219 2.49e-07 ***
## re75_cent               0.21310    0.10451   2.039 0.041883 *
## as.factor(treat)1    1381.76651  806.66779   1.713 0.087240 .
## potential_expc         48.99463   46.18898   1.061 0.289232
```

```
## potential_expc_sq      -3.71856    2.53505  -1.467 0.142937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6936 on 602 degrees of freedom
## Multiple R-squared:  0.1535, Adjusted R-squared:  0.138
## F-statistic: 9.923 on 11 and 602 DF,  p-value: < 2.2e-16
```

The r square is 0.1537 and standard error is 6941 which is not great.

Lets check the residual plots:

```
par(mfrow=c(2,2))
plot(wages.lm)
```




```
par(mfrow=c(2,2))
```

The residual plot doesn't show constant variance and qq plot is bent towards the beginning and the end and the residual does not exhibit normal distribution.

Let's buid a model by removing training and then perform an nested F-test to see if training is significant:

```
wages.lm2 <- lm(re78~ as.factor(race) + as.factor(married)  +  as.factor(educ_cat) + as.factor(race) + a
```

The r square is 0.1494 and standard error is 6947. We can see that this model has a slightly worse R squared and SE compared to the first model.

Lets do a nested F-test to check if training is significant:

```
anova(wages.lm2, wages.lm)
```

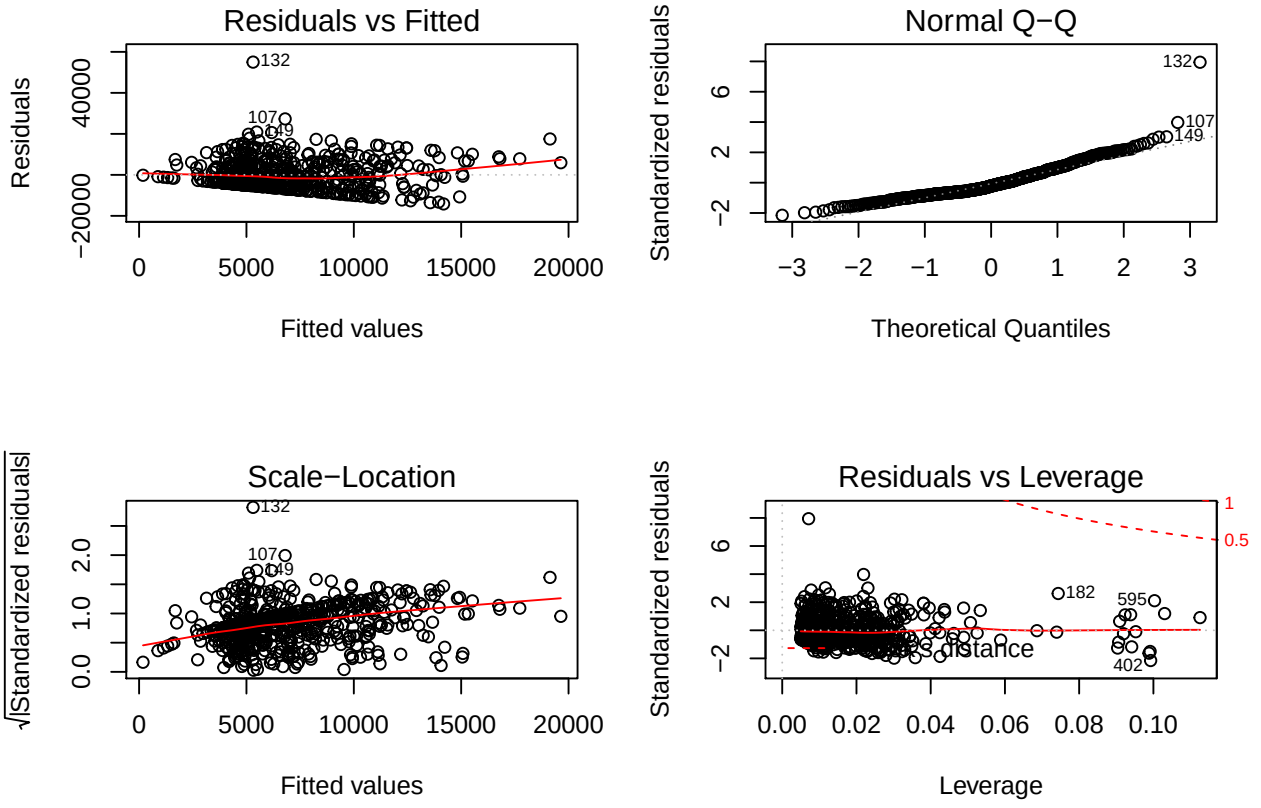
```
## Analysis of Variance Table
##
```

```
## Model 1: re78 ~ as.factor(race) + as.factor(married) + as.factor(educ_cat) +
##     as.factor(race) + as.factor(race) + re74_cent + re75_cent +
##     potential_expc + potential_expc_sq
## Model 2: re78 ~ as.factor(race) + as.factor(married) + as.factor(educ_cat) +
##     re74_cent + re75_cent + as.factor(treat) + potential_expc +
##     potential_expc_sq
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1    603 2.9103e+10
## 2    602 2.8961e+10  1 141157618 2.9341 0.08724 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the F statistics is 2.9341 and p value is 0.08724. The p value is slightly above 0.05 which suggests that training might not be significant, but we need more data to be sure.

Let's plot the residuals for the model with the interactions:

```
par(mfrow=c(2,2))
plot(wages.lm2)
```




```
par(mfrow=c(2,2))
```

We can see that the qq plot has a slight bend towards the beginning and the end of the line. This suggests that residuals do not follow normal distribution.

The residual plot does not exhibit constant variance. We can see from the residual plots that the points in the bottom are in line because the dependent variable is truncated in 0.

Nonetheless, given the observations we have and the variables available, wages.lm is the best model we can get. So let's interpret the coefficients of the model.

Confidence intervals:

```r
confint(wages.lm)
```

```
##                                2.5 %       97.5 %
## (Intercept)            1.935482e+03 6.726709e+03
## as.factor(race)2      -4.564028e+02 3.529258e+03
## as.factor(race)3      -2.134030e+02 2.800804e+03
## as.factor(married)1   -1.203397e+03 1.624779e+03
## as.factor(educ_cat)1  -4.868657e+02 3.395316e+03
## as.factor(educ_cat)2   8.875158e+02 6.110961e+03
## as.factor(educ_cat)3   1.286258e+03 1.013859e+04
## re74_cent              1.875937e-01 4.139908e-01
## re75_cent              7.850386e-03 4.183425e-01
## as.factor(treat)1     -2.024584e+02 2.965991e+03
## potential_expc        -4.171649e+01 1.397057e+02
## potential_expc_sq     -8.697184e+00 1.260063e+00
```

MODEL INTERPRETATIONS

(Results in page 3 and 4)

Intercept: For a person with average experience and avd previous annual earning (1974 and 1975) who is black and unmarried and has a middle school education level and recieved no training, the avearage value of the person's real earning in 1978 is 4331.09584 [CI: 1.935482e+03 and 6.726709e+03].

1) Is there any 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?

– Training: Keeping all the condition at the intercept (base level) constant, when the person receives training then the average value of real annual earning in 1978 increases by 1381.76651 [CI:-2.024584e+02 and 2.965991e+03].

Since the confidence interval includes a range from negative to postive values, we can conclude that job training does not have a strong effect on wages. This was also seen when we performed a nested F-test comparing model with "treat" and without "treat" variable, we found that the p value was higher than 0.05. In sum, this indicates that training is not a significant factor in determining wages in the sample we have.

2) Is there any evidence that the effects differ by demographic groups (ethnicity/education)?

Race: Keeping all the condition at the intercept constant, when the person is hispan instead of black then the average value of real annual earning in 1978 increases by 1536.42746 [CI:-4.564028e+02 and 3.529258e+03]

Keeping all the condition at the intercept constant, when the person is neither hispan nor black then the average value of real annual earning in 1978 increases by 1293.70060 [CI: -2.134030e+02 and 2.800804e+03]

Education: Keeping all the condition at the intercept constant, when the person education level changes from middle to high school then the average value of real annual earning in 1978 increases by 1454.22515 [CI: -4.868657e+02 3.395316e+03]

Keeping all the condition at the intercept constant, when the person education level changes from middle school to undergraduation then the average value of real annual earning in 1978 increases by 3499.23829 [CI: 8.875158e+02 and 6.110961e+03]

Keeping all the condition at the intercept constant, when the person education level changes from middle school to postgraduation then the average value of real annual earning in 1978 increases by 5712.42461 [CI: 1.286258e+03 and 1.013859e+04]

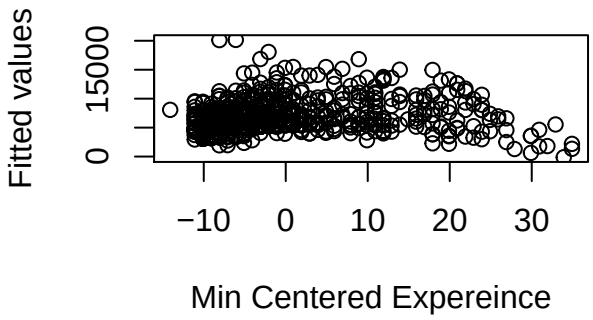
3) Are there other interesting associations with earning higher wages that are worth mentioning?

Re74 (previous earnings in 1974): Keeping all the condition at the intercept constant, when the person real annual earning in 1974 changes by 1 unit then the average value of real annual earning in 1978 changes by a multiplicative factor of 0.30079 [CI: 1.875937e-01 and 4.139908e-01]

Re75 (previous earning in 1975): Keeping all the condition at the intercept constant, when the person real annual earning in 1975 changes by 1 unit then the average value of real annual earning in 1978 changes by a multiplicative factor of 0.21310 [CI: 7.850386e-03 and 4.183425e-01]

Experience: We can undertsand the intepreation of the experience for change in real annual earning in 1978 by ploting a graph.

```
plot(y = wages.lm$fitted.values, x = lalonde$potential_expc, xlab ="Min Centered Expereince", ylab = "F
```



2) IMPACT OF TREATMENT ON POSITIVE (NON-ZERO) WAGES:

We are interested in the effect of receiving job training on non-zero wages. I will create a target variable that will be 1 if the person has positive wage and 0 if not for both times, before and after the training.

```
lalonde$posAfter <- ifelse(lalonde$re78 > 0, 1, 0)
```

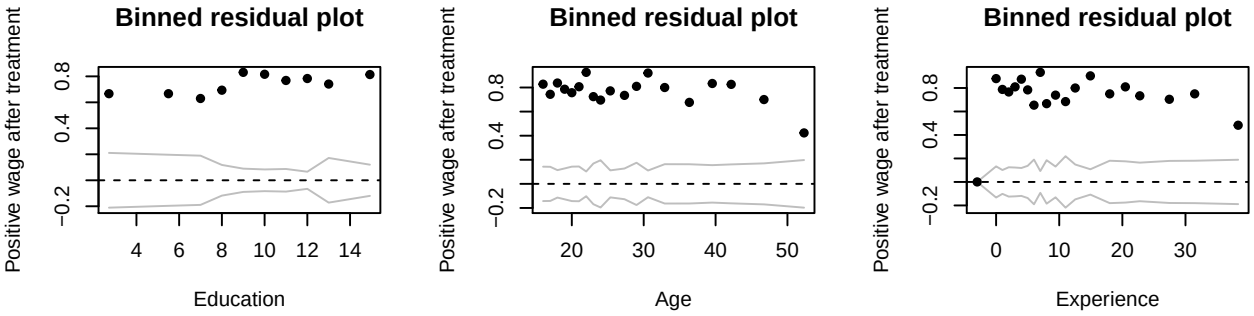
23% of the workers have non-positive wages, it's good enough to work with that data.

EXPLORATORY DATA ANALYSIS

The predictor variables available are: (1) treat: the person receive the training or not, (2) age_cent: age of the worker, (3) educ: years of education of the worker that goes from 0 to 18, (4) race: 1 if black, 2 if hipanic, 3 otherwise (5) married: 1 if married, 0 otherwise, (6) nodegree: 1 of dropped out of high school, 0 otherwise (could be highly correlated with education), (7) potential_exp: Age - years of education - 6. Assuming that immediatly after the person left the school, that person starts working, (8) educ_cat: 4 categories for education, and (9) earnings in 1974 and 1975.

Let's analyze the individual effect that the predictors has on positive wages

```
par(mfrow=c(1,3))
binnedplot(lalonde$educ, y=lalonde$posAfter, xlab = "Education", ylab = "Positive wage after treatment"
binnedplot(lalonde$age, y=lalonde$posAfter, xlab = "Age", ylab = "Positive wage after treatment")
binnedplot(lalonde$potential_exp, y=lalonde$posAfter, xlab = "Experience", ylab = "Positive wage after
```



The graph between experience and positive wage should be positive and we are observing something negative here! That suggests that I'm not getting the expected effect so I won't use the potential_exp in this analysis.

For dummy and categorical variables, after analizing the mean for each group, it seems that race and positive salary before training have an effect on positive wage. However, having a degree or not and being married or not seems to don't have a big effect on wage. Nonetheless I would say that those vars are good controls so I'll try to use it in the model. Moreover, being trated also seems to have only an small effect! Let's check the logistic regression to determine the effect of the treatment.

Finally, the earning from 1974 and 1975 have the expected positive effect we were waiting. We check the correlation between them and is aroud 0.55, nothing to worry about. We will use those previous treatment earning as predictors.

LET'S FIT THE MODEL!

```
reg1 = glm(posAfter ~ as.factor(treat) + age_cent + as.factor(educ_cat) + as.factor(race) + as.factor(ma
           + as.factor(nodegree) + re74_cent + re75_cent, data = lalonde, family = binomial)
```

Let's check if nodegree and married are significant. After computing the anova test, we found that married and nodegree are not statistically significant. However, we will keep married because it could work as a good control. We won't use nodegree because it is correlated with education as it is a function of years in high school.

We could be missing some interactions in the analysis. Based on the science behind the question, I will analyze in the combined effect of being married and receive treatment:

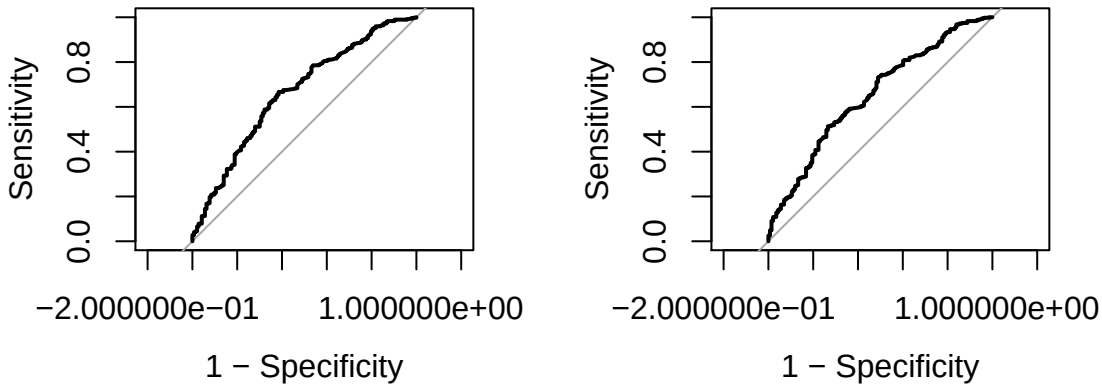
```
reg4 = glm(posAfter ~ treat + age_cent + as.factor(educ_cat) + as.factor(race) + as.factor(married) + re

reg6 = glm(posAfter ~ as.factor(treat)*as.factor(married) + age_cent + as.factor(educ_cat) + as.factor(r
```

```
par(mfrow=c(1,2))
roc(lalonde$posAfter, fitted(reg6), plot=T, legacy.axes=T)
```

```
##
## Call:
## roc.default(response = lalonde$posAfter, predictor = fitted(reg6),     plot = T, legacy.axes = T)
##
## Data: fitted(reg6) in 143 controls (lalonde$posAfter 0) < 471 cases (lalonde$posAfter 1).
## Area under the curve: 0.663
```

```
roc(lalonde$posAfter, fitted(reg4), plot=T, legacy.axes=T)
```




```
##
## Call:
## roc.default(response = lalonde$posAfter, predictor = fitted(reg4),     plot = T, legacy.axes = T)
##
## Data: fitted(reg4) in 143 controls (lalonde$posAfter 0) < 471 cases (lalonde$posAfter 1).
## Area under the curve: 0.653
```

After looking at the anova test, we found that the interaction between treatment and being married seems to be quite significant, the are under the curve looks good but pretty similar than the regression without the interaction. The interpretation of the coefficients as well as the likely range for the effect of training is really difficult when we employ interactions. Thus, we will keep the simplest model (without interaction) since the performance is really similar between both.

Now, since the main effect we are looking at is the effect of treatment on positive wages, we will employ an anova test to identify if receiving the training is significant or not:

```
reg4_notreat = glm(posAfter ~ age_cent + as.factor(educ_cat) + as.factor(race) + as.factor(married) + re

anova(reg4, reg4_notreat, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: posAfter ~ treat + age_cent + as.factor(educ_cat) + as.factor(race) +
##     as.factor(married) + re74_cent + re75_cent
## Model 2: posAfter ~ age_cent + as.factor(educ_cat) + as.factor(race) +
##     as.factor(married) + re74_cent + re75_cent
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       603     628.75
## 2       604     630.61 -1  -1.8616   0.1724
```

Seems that receiving training does not have an influence on wages. However, we will keep the treat variable in the model as a control to interpret the other explanatory variables.

Let's do the binned residuals plots of this final model:

```
par(mfrow=c(1,3))
rawresid4 = lalonde$posAfter - fitted(reg4)
binnedplot(x=lalonde$age_cent, y = rawresid4, xlab = "Age centered", ylab = "Residuals", main = "Binned
binnedplot(x=lalonde$re74_cent, y = rawresid4, xlab = "Earnings 1974 centered", ylab = "Residuals", mai
binnedplot(x=lalonde$re75_cent, y = rawresid4, xlab = "Earnings 1975 centered", ylab = "Residuals", mai
```

Seems that the residuals vs. age are randomly distributed! There is just a little trouble with the earnings from 1975 for the highest earning, but that could be because we don't have a lot of people with such big earnings.

```
tapply(rawresid4, lalonde$treat, mean)
```

```
##            0            1
## 4.285315e-10 3.065475e-10
```

```
tapply(rawresid4, lalonde$married, mean)
```

```
##            0            1
## 1.478730e-10 7.351564e-10
```

```
tapply(rawresid4, lalonde$educ_cat, mean)
```

```
##            0            1            2            3
## 3.800131e-10 3.926459e-10 3.797871e-10 4.861530e-10
```

```
tapply(rawresid4, lalonde$race, mean)
```

```
##            1            2            3
## 3.642116e-10 4.622319e-10 3.972147e-10
```

We are expecting to see numbers near to zero and that's what we are getting! So the model seems to fit the data pretty well.

Let's do the confusion matrix

```
threshold = 0.50
table(lalonde$posAfter, reg4$fitted > threshold)
```

```
##
##     FALSE TRUE
##   0    10  133
##   1     8  463
```

```
Accuracy = (14+466)/(14+5+129+466)
Precision = (466)/(129+466)
Recall = (466)/(466+5)
```

The numbers in the diagonal (true positives and true negatives) seems good. The accuracy=0.78, precision=0.78, and recall=0.99. We fell comfortable being able to predict those percentages of successes.

MODEL INTERPRETATIONS

```
summary(reg4)
```

```
##
## Call:
## glm(formula = posAfter ~ treat + age_cent + as.factor(educ_cat) +
##     as.factor(race) + as.factor(married) + re74_cent + re75_cent,
##     family = binomial, data = lalonde)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3219   0.3668   0.6247   0.7626   1.4839
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           4.477e-01  3.501e-01   1.279  0.20089
## treat                 3.665e-01  2.696e-01   1.359  0.17403
## age_cent             -3.524e-02  1.076e-02  -3.276  0.00105 **
## as.factor(educ_cat)1  4.115e-01  2.997e-01   1.373  0.16976
## as.factor(educ_cat)2  5.252e-01  4.321e-01   1.215  0.22420
## as.factor(educ_cat)3  2.731e-01  7.434e-01   0.367  0.71331
## as.factor(race)2      7.466e-01  3.824e-01   1.952  0.05091 .
## as.factor(race)3      5.439e-01  2.639e-01   2.061  0.03932 *
## as.factor(married)1   1.102e-02  2.418e-01   0.046  0.96366
## re74_cent             3.555e-05  2.188e-05   1.625  0.10427
## re75_cent             9.627e-05  4.557e-05   2.113  0.03462 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 666.50  on 613  degrees of freedom
## Residual deviance: 628.75  on 603  degrees of freedom
## AIC: 650.75
##
## Number of Fisher Scoring iterations: 4
```

```
exp(confint.default(reg4))
```

```
##                           2.5 %    97.5 %
## (Intercept)           0.7879082 3.1075810
## treat                 0.8505025 2.4472058
## age_cent              0.9452288 0.9859401
## as.factor(educ_cat)1 0.8386852 2.7152379
## as.factor(educ_cat)2 0.7249202 3.9432973
## as.factor(educ_cat)3 0.3060686 5.6419933
## as.factor(race)2      0.9970512 4.4644908
## as.factor(race)3      1.0269623 2.8895708
## as.factor(married)1   0.6294875 1.6239810
## re74_cent             0.9999927 1.0000784
## re75_cent             1.0000070 1.0001856
```

1) Is there any 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?

– For a person with average age, average previous earnings (1975 and 1974), not married, black race, and with middle school education, having the treatment will change the odds of having a positive wage by a multiplicative factor of exp(0.3665)=1.56. The likely range of this change is (0.85 - 2.44).

However, as we saw in the anova test above, training is not a significant variable.

2) Is there any evidence that the effects differ by demographic groups (ethnicity/education)?

– RACE: For a person with average age, average previous earnings (1975 and 1974), not married, with middle school education, and with treatment, being hispanic instead of black (baseline), will change the odds of having a positive wage by a multiplicative factor of exp(0.0.7466)=2.1098 with a CI = (0.997 - 4.464). And for a person with the same characteristics mentioned above, being other race instead of black (baseline) the odds of having a positive wage will change by a multiplicative factor of exp(0.0.7466)=2.1098 with a CI = (0.997 - 4.464).

– EDUCATION: For a person with average age, average previous earnings (1975 and 1974), not married, black race, and with treatment, going from middle school education (baseline) to high school, represents a change in the odds of having a positive wage by a multiplicative factor of exp(0.41)=1.51, CI = (0.83 - 2.72); when going from middle school to undergraduation the change will be exp(0.52)=1.69, CI = (0.72 - 3.94); and when going from middle school to postgraduation will be exp(0.27)=1.31, CI = (0.31 - 5.64). The odds of having a positive wage increase when the education attained increase. Except for the case of postgraduation. This can be happening because we do not have enough information of people with postgraduate education.

3) Are there other interesting associations with positive wages that are worth mentioning?

– AGE: For a person with average previous earnings (1975 and 1974), not married, black race, in the treatment group, and with middle school education, having the treatment will change the odds of having a positive wage by a multiplicative factor of exp(-0.3665)=0.69, CI = (0.95 - 0.99). This means that when the worker gets older, the probability of having a wage greater than zero decrease.

CONCLUSIONS AND LIMITATIONS:

a) The R-Square in the linear model is not very high. Thus, this model can not be a definitive conclusion on the factors affecting earnings.

b) Also, the anova test in the logistic regression shows that receive the treatment is not significant to predict if the wage will be zero or greater than zero.

c) The wages data is truncated in zero. It would be a better approach to use a "Two-Part Model" instead of a linear regression and then a logistic regression without deleting wages=0. A more sophisticated model could be more appropriate to address the questions.

d) A possible limitation related with the data available, is that we do not have information about the real work experience of the workers. We build a proxy variable (potential experience) but it has been shown in the literature that is not as good as the real experience when determining wages.

e) Moreover, in the data available we do not have information of the IQ of the workers. It has been shown in the labor economics literature that the wages can vary depending on the IQ of the people and we are not controlling this effect in the models.