Team Project - Effects of Job Training on Wages

Sicong Zhao, Xuan Yu

10/18/2018

Question2

```
library("pROC")
library("arm")
```

First we read in the data, take a first look and then do the mean centering:

```
ldata <- read.table('lalondedata', header = TRUE, sep = ',')
dim(ldata)
```

```
## [1] 614  11
```

```
summary(ldata)
```

```
##        X            treat             age             educ      
##  NSW1   :  1   Min.   :0.0000   Min.   :16.00   Min.   : 0.00  
##  NSW10  :  1   1st Qu.:0.0000   1st Qu.:20.00   1st Qu.: 9.00  
##  NSW100 :  1   Median :0.0000   Median :25.00   Median :11.00  
##  NSW101 :  1   Mean   :0.3013   Mean   :27.36   Mean   :10.27  
##  NSW102 :  1   3rd Qu.:1.0000   3rd Qu.:32.00   3rd Qu.:12.00  
##  NSW103 :  1   Max.   :1.0000   Max.   :55.00   Max.   :18.00  
##  (Other):608                                                   
##      black            hispan           married          nodegree     
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :1.0000  
##  Mean   :0.3958   Mean   :0.1173   Mean   :0.4153   Mean   :0.6303  
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##                                                                     
##       re74             re75              re78        
##  Min.   :    0   Min.   :    0.0   Min.   :    0.0  
##  1st Qu.:    0   1st Qu.:    0.0   1st Qu.:  238.3  
##  Median : 1042   Median :  601.5   Median : 4759.0  
##  Mean   : 4558   Mean   : 2184.9   Mean   : 6792.8  
##  3rd Qu.: 7888   3rd Qu.: 3249.0   3rd Qu.:10893.6  
##  Max.   :35040   Max.   :25142.2   Max.   :60307.9  
##                                                     
```

```
ldata$age.c = ldata$age - mean(ldata$age)
ldata$educ.c = ldata$educ - mean(ldata$educ)
```

We started from making a dummy variable to show if salary is positive in 1978, and creating two dummy varible as predictor variable: employed_74 and employed_75

```
ldata$positive_sal78 = rep(0, nrow(ldata))
ldata$positive_sal78[ldata$re78 != 0] = 1
ldata$employed_74 = rep(0, nrow(ldata))
ldata$employed_75 = rep(0, nrow(ldata))
```

```
ldata$employed_74[ldata$re74 != 0] = 1
ldata$employed_75[ldata$re75 != 0] = 1
```

Exploratory data analysis:

See Appendix for details.

After some Exploratory data analysis, we then try a logistic regression that has a main effect for every variable and linear predictors:

```
q2reg1 = glm(positive_sal78 ~ age.c + educ.c + re74 + re75 + as.factor(employed_74) + as.factor(employe
summary(q2reg1)
```

```
##
## Call:
## glm(formula = positive_sal78 ~ age.c + educ.c + re74 + re75 +
##     as.factor(employed_74) + as.factor(employed_75) + as.factor(black) +
##     as.factor(hispan) + as.factor(married) + as.factor(nodegree) +
##     as.factor(treat), family = binomial, data = ldata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2941   0.3665   0.6260   0.7641   1.4589
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             9.141e-01  3.191e-01   2.865 0.004176 **
## age.c                  -3.640e-02  1.093e-02  -3.329 0.000871 ***
## educ.c                  4.804e-02  5.346e-02   0.899 0.368828
## re74                    3.926e-05  2.423e-05   1.620 0.105147
## re75                    9.221e-05  5.204e-05   1.772 0.076445 .
## as.factor(employed_74)1 -8.309e-02  2.834e-01  -0.293 0.769372
## as.factor(employed_75)1  6.602e-02  2.692e-01   0.245 0.806292
## as.factor(black)1      -5.356e-01  2.659e-01  -2.014 0.044020 *
## as.factor(hispan)1      2.138e-01  3.643e-01   0.587 0.557325
## as.factor(married)1     1.624e-02  2.412e-01   0.067 0.946305
## as.factor(nodegree)1    1.081e-01  2.986e-01   0.362 0.717449
## as.factor(treat)1       3.722e-01  2.787e-01   1.336 0.181598
## ---
## 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: 629.76  on 602  degrees of freedom
## AIC: 653.76
##
## Number of Fisher Scoring iterations: 4
```

Model diagnostics

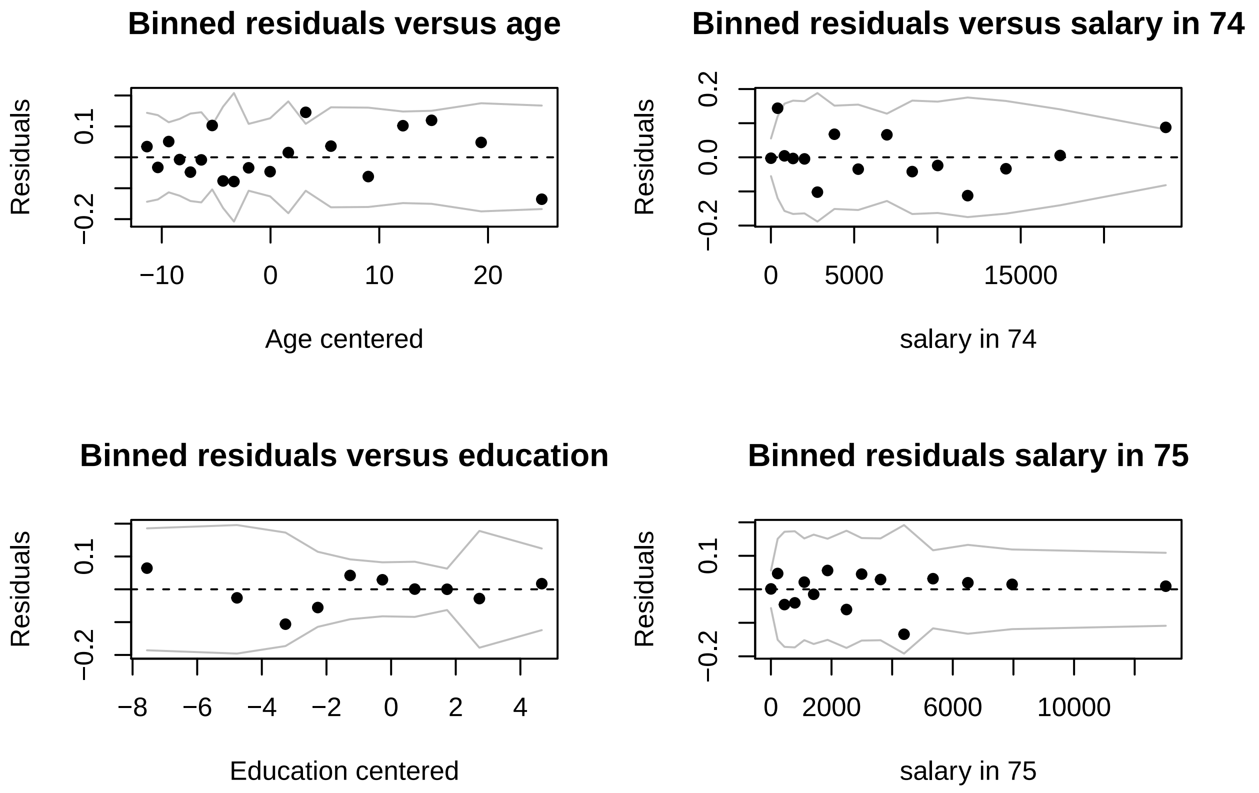
We first do binned residual plots for numeric variables: We don't find any noticeable patterns.

```r
par(mfcol = c(2,2))
rawresid1 = ldata$positive_sal78 - fitted(q2reg1)
binnedplot(x=ldata$age.c, y = rawresid1, xlab = "Age centered", ylab = "Residuals",
main = "Binned residuals versus age")

binnedplot(x=ldata$educ.c, y = rawresid1, xlab = "Education centered", ylab = "Residuals",
main = "Binned residuals versus education")

binnedplot(x=ldata$re74, y = rawresid1, xlab = "salary in 74", ylab = "Residuals",
main = "Binned residuals versus salary in 74")

binnedplot(x=ldata$re75, y = rawresid1, xlab = "salary in 75", ylab = "Residuals",
main = "Binned residuals salary in 75")
```



Then look at average residuals by dummy variables using the tapply command: Nothing specific.

```r
tapply(rawresid1, ldata$black, mean)
```

```
##            0            1
## 1.029411e-09 8.307175e-10
```

```r
tapply(rawresid1, ldata$hispan, mean)
```

```
##            0            1
## 9.315583e-10 1.095431e-09
```

```r
tapply(rawresid1, ldata$married, mean)
```

```
##            0            1
## 2.972536e-10 1.870830e-09
```

```
tapply(rawresid1, ldata$nodegree, mean)
```

```
##            0            1
## 9.866986e-10 9.297030e-10
```

```
tapply(rawresid1, ldata$treat, mean)
```

```
##            0            1
## 1.073914e-09 6.652248e-10
```

```
tapply(rawresid1, ldata$employed_74, mean)
```

```
##            0            1
## 1.008136e-10 1.507488e-09
```

```
tapply(rawresid1, ldata$employed_75, mean)
```

```
##            0            1
## 6.417200e-12 1.577787e-09
```

Confusion matrix with .5 threshold and .6 threshold:

```
threshold = 0.6
table(ldata$positive_sal78, q2reg1$fitted > threshold)
```

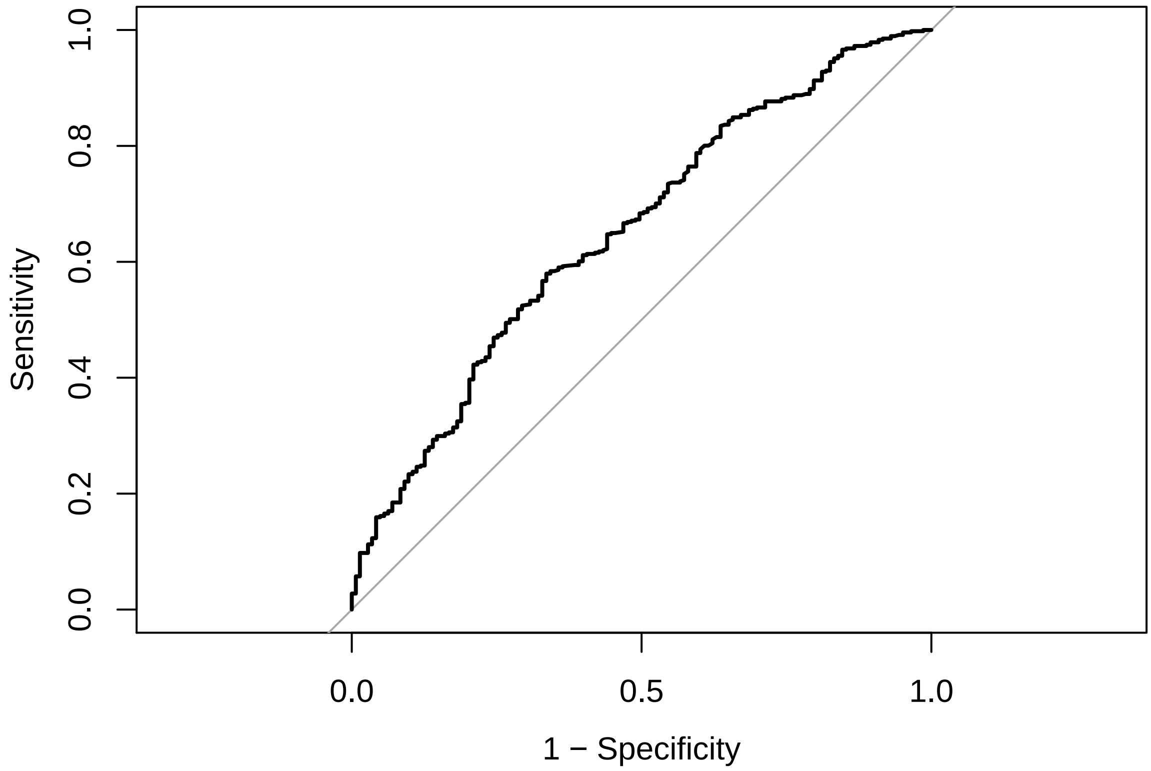
```
##
##     FALSE TRUE
##   0    22  121
##   1    19  452
```

```
threshold = 0.5
table(ldata$positive_sal78, q2reg1$fitted > threshold)
```

```
##
##     FALSE TRUE
##   0    10  133
##   1     6  465
```

Then we look at ROC curve: We didn't find specific pattern from the model diagnostics, so we decide not to do transformations. We got area under the curve value: 0.6501.

```
roc(ldata$positive_sal78, fitted(q2reg1), plot=T, legacy.axes=T)
```



```
##
## Call:
## roc.default(response = ldata$positive_sal78, predictor = fitted(q2reg1),     plot = T, legacy.axes =
##
## Data: fitted(q2reg1) in 143 controls (ldata$positive_sal78 0) < 471 cases (ldata$positive_sal78 1).
## Area under the curve: 0.6501
```

We then need to then check if the effects differ by demographic groups. We first look at the p value of the original model and find Age and Black variable may have a protential impact on the outcome.

So we first try the model without Age and Black variable:

```
q2reg2 = glm(positive_sal78 ~ educ.c + re74 + re75 + as.factor(employed_74) + as.factor(employed_75) +
```

We do the change in deviance test to see if the effects differ by these two variables. We get a p value of 0.0006713, so we find the effect differ by age and race black.

```
anova(q2reg1, q2reg2, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: positive_sal78 ~ age.c + educ.c + re74 + re75 + as.factor(employed_74) +
##     as.factor(employed_75) + as.factor(black) + as.factor(hispan) +
##     as.factor(married) + as.factor(nodegree) + as.factor(treat)
## Model 2: positive_sal78 ~ educ.c + re74 + re75 + as.factor(employed_74) +
##     as.factor(employed_75) + as.factor(hispan) + as.factor(married) +
##     as.factor(nodegree) + as.factor(treat)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       602     629.76
## 2       604     644.37 -2  -14.613 0.0006713 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And we're also interested in education level and want to make sure whether the effect changes by demographic groups of different education level, so we remove Education variable and do the change in deviance test, and we get a p value of 0.3695, so education level might not be so important, but we will keep it in the model:

```
q2reg3 = glm(positive_sal78 ~ age.c + re74 + re75 + as.factor(employed_74) + as.factor(employed_75) + as
anova(q2reg1, q2reg3, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: positive_sal78 ~ age.c + educ.c + re74 + re75 + as.factor(employed_74) +
##     as.factor(employed_75) + as.factor(black) + as.factor(hispan) +
##     as.factor(married) + as.factor(nodegree) + as.factor(treat)
## Model 2: positive_sal78 ~ age.c + re74 + re75 + as.factor(employed_74) +
##     as.factor(employed_75) + as.factor(black) + as.factor(hispan) +
##     as.factor(married) + as.factor(nodegree) + as.factor(treat)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       602     629.76
## 2       603     630.56 -1 -0.80539   0.3695
```

Then, scientifically, it is plausible to think that there might be interactions between age variable and nodegree and treat variable, because older people might focus less on their studying and might have a hard time learning new things even they received a job training, which leads to greater odds of having 0 income than younger people; And older people with no degree may be way more hard to find a job than younger people with no degree. So I might try the interaction between age variable and nodegree and treat variable:

```
q2reg4 = glm(positive_sal78 ~ age.c * (as.factor(nodegree) + as.factor(treat)) + educ.c + re74 + re75 +
```

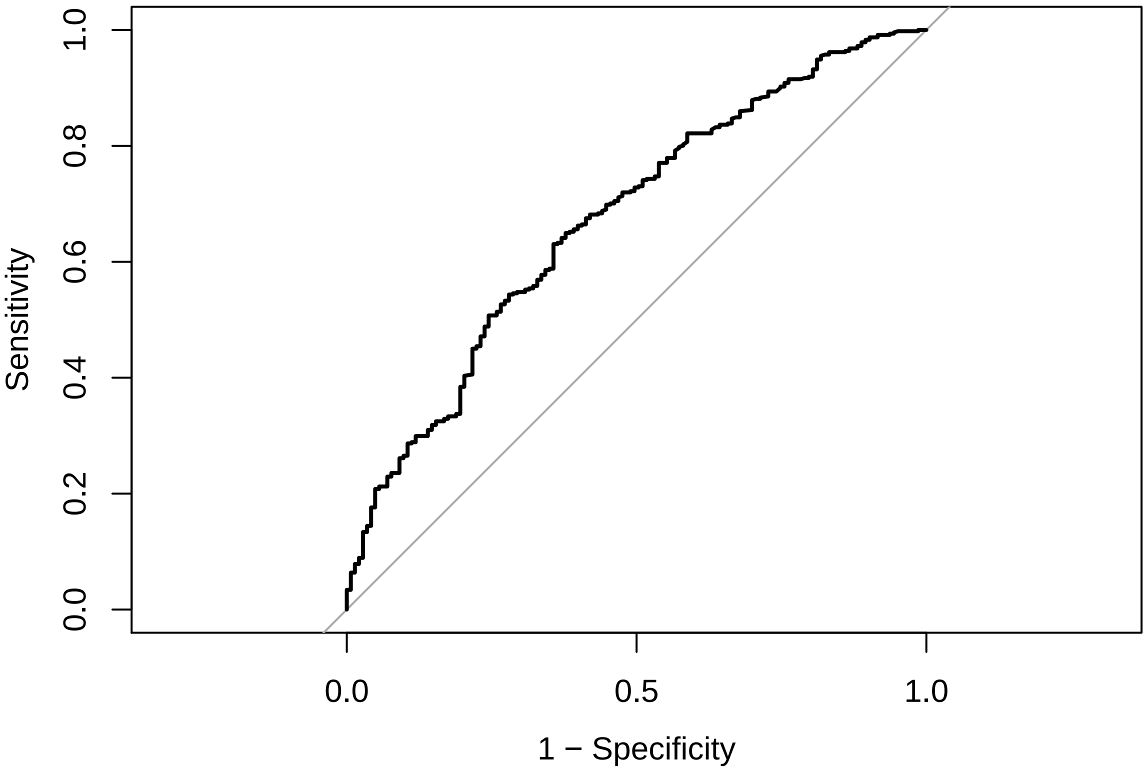
We do the change in deviance test to see if the interaction is useful. We get a p value of 0.002642, so we might find interaction between age variable and nodegree and treat variable.

```
anova(q2reg1, q2reg4, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: positive_sal78 ~ age.c + educ.c + re74 + re75 + as.factor(employed_74) +
##     as.factor(employed_75) + as.factor(black) + as.factor(hispan) +
##     as.factor(married) + as.factor(nodegree) + as.factor(treat)
## Model 2: positive_sal78 ~ age.c * (as.factor(nodegree) + as.factor(treat)) +
##     educ.c + re74 + re75 + as.factor(employed_74) + as.factor(employed_75) +
##     as.factor(black) + as.factor(hispan) + as.factor(married)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       602     629.76
## 2       600     617.88  2   11.872 0.002642 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then we check the ROC curve of this new model: We got a lightly better area under the curve value: 0.6734.

```
roc(ldata$positive_sal78, fitted(q2reg4), plot=T, legacy.axes=T)
```



```
##
## Call:
## roc.default(response = ldata$positive_sal78, predictor = fitted(q2reg4),     plot = T, legacy.axes =
##
## Data: fitted(q2reg4) in 143 controls (ldata$positive_sal78 0) < 471 cases (ldata$positive_sal78 1).
## Area under the curve: 0.6734
```

Because the interaction we found is scientifically reasonable, and we got a way better ROC curve with that model, we decide to choose model_4 as our final model. Here is the model:

```
summary(q2reg4)
```

```
##
## Call:
## glm(formula = positive_sal78 ~ age.c * (as.factor(nodegree) +
##     as.factor(treat)) + educ.c + re74 + re75 + as.factor(employed_74) +
##     as.factor(employed_75) + as.factor(black) + as.factor(hispan) +
##     as.factor(married), family = binomial, data = ldata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3186   0.3429   0.5892   0.7561   1.6789
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  8.221e-01  3.224e-01   2.550  0.01078 *
## age.c                       -1.689e-02  1.936e-02  -0.873  0.38285
## as.factor(nodegree)1         7.900e-02  2.991e-01   0.264  0.79164
## as.factor(treat)1            5.005e-01  2.878e-01   1.739  0.08202 .
## educ.c                       1.940e-02  5.512e-02   0.352  0.72493
## re74                         4.355e-05  2.483e-05   1.754  0.07951 .
## re75                         9.664e-05  5.343e-05   1.809  0.07047 .
## as.factor(employed_74)1      3.520e-02  2.909e-01   0.121  0.90370
## as.factor(employed_75)1      7.465e-02  2.731e-01   0.273  0.78459
## as.factor(black)1           -5.141e-01  2.718e-01  -1.892  0.05854 .
## as.factor(hispan)1           2.043e-01  3.684e-01   0.555  0.57916
## as.factor(married)1          1.356e-02  2.471e-01   0.055  0.95622
## age.c:as.factor(nodegree)1 -4.852e-02  2.212e-02  -2.194  0.02823 *
## age.c:as.factor(treat)1      7.538e-02  2.787e-02   2.704  0.00684 **
## ---
## 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: 617.88  on 600  degrees of freedom
## AIC: 645.88
##
## Number of Fisher Scoring iterations: 4
```

And we take exponential for the coefficients and confidence intervals to get the odds:

```r
exp(q2reg4$coefficients)
```

```
##                (Intercept)                      age.c
##                  2.2752006                  0.9832486
##        as.factor(nodegree)1          as.factor(treat)1
##                  1.0822081                  1.6495654
##                     educ.c                       re74
##                  1.0195846                  1.0000435
##                       re75    as.factor(employed_74)1
##                  1.0000966                  1.0358244
##    as.factor(employed_75)1          as.factor(black)1
##                  1.0775123                  0.5980586
##         as.factor(hispan)1        as.factor(married)1
##                  1.2267122                  1.0136565
## age.c:as.factor(nodegree)1    age.c:as.factor(treat)1
##                  0.9526351                  1.0782960
```

```r
exp(confint(q2reg4))
```

```
##                                2.5 %    97.5 %
## (Intercept)                1.2157472 4.3121422
## age.c                      0.9471404 1.0223448
## as.factor(nodegree)1       0.6013966 1.9459076
## as.factor(treat)1          0.9423436 2.9186216
## educ.c                     0.9145179 1.1357889
## re74                       0.9999961 1.0000938
## re75                       0.9999970 1.0002070
## as.factor(employed_74)1    0.5848197 1.8334640
## as.factor(employed_75)1    0.6306635 1.8431476
## as.factor(black)1          0.3501815 1.0182215
## as.factor(hispan)1         0.6118956 2.6207754
## as.factor(married)1        0.6262841 1.6530981
## age.c:as.factor(nodegree)1 0.9113761 0.9942318
## age.c:as.factor(treat)1    1.0224624 1.1413353
```

Interpretation:

Holding other variables constant, in the average age of workers, when shifting from not taking the job training to taking the training, the odds of workers having positive salary tend to be 164.96% higher than before, with a 95% confidence interval of (94.23, 291.86).

Also, there is evidence that the effects differ by demographic groups of different ages and black people. We did change in deviance test and got a p value of 0.0006713. Worker of different ages or their race are black or not may lead to different odds of getting positive salary.

I also found several interesting associations with the odds of getting positive salary:

1, Whether workers' race are black or not really leads to a significant difference to the odds of them having positive salary in year 78: holding other variables constant, the odds of black workers having positive salary tend to be 59.81% of non-black workers, with a 95% confidence interval of (35.02%, 101.82%).

2, Due to the p value, we found that workers' salary in year 74 and 75 might have some effect on whether they are getting positive salary or not. But the interesting part is, these two variable both have exponentiated coefficients of nearly 100%, which means that the difference of workers' salary in both year 74 and 75 not really changes the odds of them getting positive salary in year 78: e.g. holding other variables constant, when workers' salary in year 74 doubles, the odd of them having a positive salary in year 78 will be multiply by exp(4.355e-05 * log2) = 100.003%, which is very interesting.

Limitations:

1, Treat variable has a large standard error and has a confidence interval which contains odds below 100%. So even we might find an impact of training for workers to get jobs in year 78 and it seems to be positive, we cannot say we are 95% percent sure it does have positive impact. We need larger data set to confirm that.

2, Black and Hispanic are both race related variables, but Black seems to have effect on the result while Hispanic not. We need to ask some experts about the reason and make sure our conclusion is reasonable.