

Team Project - Effects of Job Training on Wages

Sicong Zhao, Xuan Yu

10/18/2018

Question2

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

```
library("pROC")
library("arm")
ldata <- read.table('lalondedata', header = TRUE, sep = ',')
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 variable 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
```

After some Exploratory data analysis, we then try a logistic regression that has a main effect for every variable and linear predictors, we put the summary of the model and all diagnostics part in the appendix, and we got area under the curve value 0.6501:

```
q2reg1 = glm(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),
             data = ldata, family = binomial)
```

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 potential 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) + as.factor(hispan) + as.factor(married) +
             as.factor(nodegree) + as.factor(treat),
             data = ldata, family = binomial)
```

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
```

We're interested in education level, so we build a model 3. We remove Education variable and do the change in deviance test, and put the analytics part in the appendix. We get a p value of 0.3695, so education level might not be so important, but we will keep it in the model.

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 + as.factor(employed_74) + as.factor(employed_75) +
            as.factor(black) + as.factor(hispan) + as.factor(married),
            data = ldata, family = binomial)
```

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
```

Model diagnostics

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

```
par(mfcol = c(2,2))
rawresid4 = ldata$positive_sal78 - fitted(q2reg4)
binnedplot(x=ldata$age.c, y = rawresid4, xlab = "Age centered", ylab = "Residuals",
main = "Binned residuals versus age")

binnedplot(x=ldata$educ.c, y = rawresid4, xlab = "Education centered", ylab = "Residuals",
main = "Binned residuals versus education")

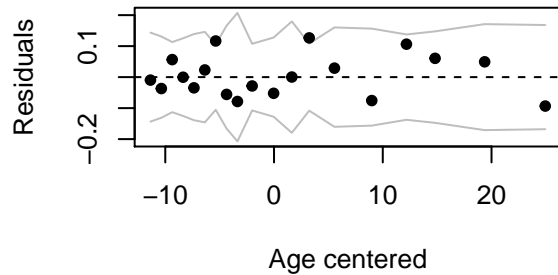
binnedplot(x=ldata$re74, y = rawresid4, xlab = "salary in 74", ylab = "Residuals",
main = "Binned residuals versus salary in 74")
```

```

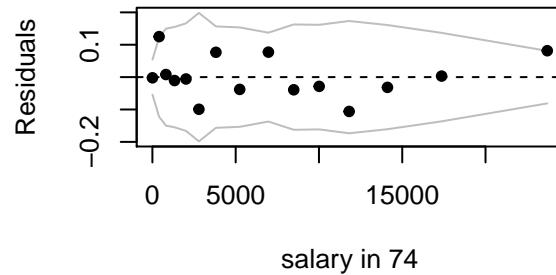
binnedplot(x=ldata$re75, y = rawresid4, xlab = "salary in 75", ylab = "Residuals",
main = "Binned residuals salary in 75")

```

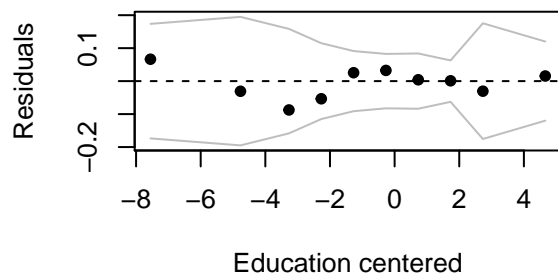
Binned residuals versus age



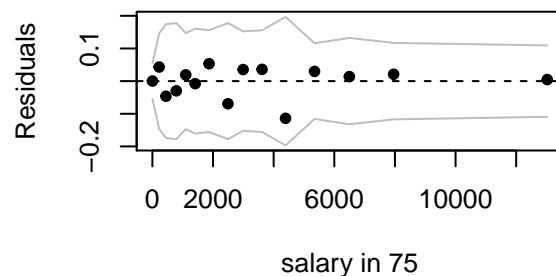
Binned residuals versus salary in 74



Binned residuals versus education



Binned residuals salary in 75



Then look at average residuals by dummy variables using the `tapply` command:

```

tapply(rawresid4, ldata$black, mean);tapply(rawresid4, ldata$hispan, mean);tapply(rawresid4, ldata$marr

```

```

##           0           1
## 3.206778e-09 2.204458e-09
##           0           1
## 2.731825e-09 3.399292e-09
##           0           1
## 9.225977e-10 5.467393e-09
##           0           1
## 3.140466e-09 2.616310e-09
##           0           1
## 3.293155e-09 1.689916e-09
##           0           1
## 2.898782e-10 4.460802e-09
##           0           1
## -1.669067e-13 4.675986e-09

```

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

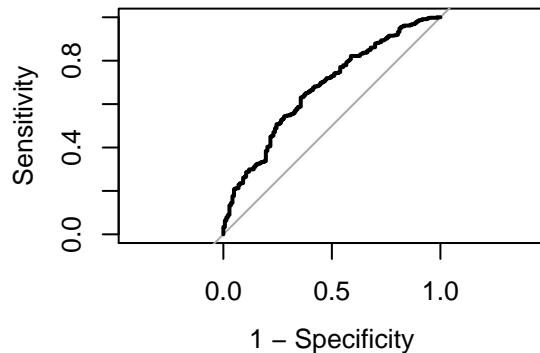
```

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



Choosing the model:

Because the interactions we found is scientifically reasonable, and all residual plots seem fine; we also got a way better ROC curve with the model, we decide to choose model 4 as our final model.

Here is the summary of the final 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
```

For the interpretation, we take exponential for the confidence intervals to get the odds:

```
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 and Answers:

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 $\exp(4.355e-05 * \log 2) = 100.003\%$, 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.