STOR 455 Homework 7

25 points - Due Friday 4/1 9:00am

Are Emily and Greg More Employable Than Lakisha and Jamal?

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), pp. 991-1013.

Abstract

We perform a field experiment to measure racial discrimination in the labor market. We respond with fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perception of race, each resume is randomly assigned either a very African American sounding name or a very White sounding name. The results show significant discrimination against African-American names: White names receive 50 percent more callbacks for interviews. We also find that race affects the benefits of a better resume. For White names, a higher quality resume elicits 30 percent more callbacks whereas for African Americans, it elicits a far smaller increase. Applicants living in better neighborhoods receive more callbacks but, interestingly, this effect does not differ by race. The amount of discrimination is uniform across occupations and industries. Federal contractors and employers who list "Equal Opportunity Employer" in their ad discriminate as much as other employers. We find little evidence that our results are driven by employers inferring something other than race, such as social class, from the names. These results suggest that racial discrimination is still a prominent feature of the labor market.

Variables	Descriptions
call	Was the applicant called back? (1 = yes; 0 = no)
ethnicity	indicating ethnicity (i.e., "Caucasian-sounding" vs. "African-American sounding" first name)
sex	indicating sex
quality	Indicating quality of resume.
experience	Number of years of work experience on the resume
equal	Is the employer EOE (equal opportunity employment)?

Use the *ResumeNames455* found at the address below:

https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/ResumeNames455.csv

```r
library(readr)
ResumeNames455 <- read_csv("https://raw.githubusercontent.com/JA-McLean
/STOR455/master/data/ResumeNames455.csv")

## Rows: 4870 Columns: 7

## -- Column specification --------------------------------------------
------------
## Delimiter: ","
## chr (5): name, sex, ethnicity, quality, equal
## dbl (2): call, experience

##
## i Use `spec()` to retrieve the full column specification for this da
ta.
## i Specify the column types or set `show_col_types = FALSE` to quiet
this message.
```

1) An Equal Opportunity Employer (EOE) is an employer who agrees not to discriminate against any employee or job applicant because of race, color, religion, national origin, sex, physical or mental disability, or age. Construct a logistic model to predict if the job applicant was called back using *ethnicity*, *equal*, *sex*, and the interactions between *ethnicity* and *equal*, and *sex* and *equal* as the predictor variables.

```r
mod=glm(call~ethnicity+sex+equal+ethnicity*equal+sex*equal,family = bin
omial, data=ResumeNames455)
summary(mod)

##
## Call:
## glm(formula = call ~ ethnicity + sex + equal + ethnicity * equal +
##     sex * equal, family = binomial, data = ResumeNames455)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4990  -0.4570  -0.3803  -0.3453   2.4241
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.5905     0.1012 -25.590  < 2e-16 ***
## ethnicitycauc              0.3839     0.1274   3.014  0.00258 **
## sexmale                   -0.2932     0.1621  -1.809  0.07052 .
## equalyes                  -0.1997     0.1960  -1.019  0.30835
## ethnicitycauc:equalyes     0.1877     0.2372   0.791  0.42878
## sexmale:equalyes           0.4910     0.2704   1.816  0.06941 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 2726.9  on 4869  degrees of freedom
## Residual deviance: 2705.0  on 4864  degrees of freedom
## AIC: 2717
##
## Number of Fisher Scoring iterations: 5
```

2) Conduct a drop in deviance hypothesis test to determine the effectiveness of the *equal* terms in the model constructed in the previous question. Cite your hypotheses, p-value, and conclusion in context. H0:β3=0 Ha:β3≠0 We cannot reject H0(p-value=0.2629), the equal term cannot significantly improve the model.

```
mod2=glm(call~ethnicity+sex,family = binomial, data=ResumeNames455)
anova(mod2, mod, test="Chisq")

## Analysis of Deviance Table
##
## Model 1: call ~ ethnicity + sex
## Model 2: call ~ ethnicity + sex + equal + ethnicity * equal + sex *
equal
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4867     2708.9
## 2      4864     2705.0  3   3.9869   0.2629

1 - pchisq(summary(mod2)$deviance - summary(mod)$deviance, 3)

## [1] 0.2628813
```

3) Based on your model from question 1, What is the probability of a male applicant with a "Caucasian-sounding" name getting a call back from an Equal Opportunity Employer (EOE)? What is the probability of a female applicant with an "African-American sounding" name getting a call back from an Equal Opportunity Employer (EOE)? Person1 11.70% person2 5.79%

```
person1=data.frame(sex='male',ethnicity='cauc',equal='yes')
person2=data.frame(sex='female',ethnicity='afam',equal='yes')
predict(mod,person1, type = "response")

##         1
## 0.1170428

predict(mod,person2, type = "response")

##          1
## 0.05785914
```

4) Does the number of years of work experience impact the relationship between *ethnicity*, *sex*, and an applicant getting called back? Construct a logistic model to predict if the job applicant was called back using *ethnicity*, *sex*, *experience*, and the interactions between *ethnicity* and *experience*, and *sex* and *experience* as the predictor variables.

```
mod3=glm(call~ethnicity+sex+experience+ethnicity*experience+sex*experie
nce,family = binomial, data=ResumeNames455)
summary(mod3)

##
## Call:
## glm(formula = call ~ ethnicity + sex + experience + ethnicity *
##     experience + sex * experience, family = binomial, data = ResumeN
ames455)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8867  -0.4320  -0.3941  -0.3458   2.4913
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -3.111554   0.159022 -19.567  < 2e-16 ***
## ethnicitycauc             0.497435   0.196458   2.532  0.01134 *
## sexmale                   0.351931   0.230841   1.525  0.12737
## experience                0.054109   0.014646   3.694  0.00022 ***
## ethnicitycauc:experience -0.006006   0.018719  -0.321  0.74831
## sexmale:experience       -0.057080   0.024796  -2.302  0.02133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2726.9  on 4869  degrees of freedom
## Residual deviance: 2686.4  on 4864  degrees of freedom
## AIC: 2698.4
##
## Number of Fisher Scoring iterations: 5
```

5) Conduct a drop in deviance hypothesis test to determine the effectiveness of the *experience* terms in the model constructed in the previous question. Cite your hypotheses, p-value, and conclusion in context. H0:β3=0 Ha:β3≠0 We can reject H0(p-value=4.958e-05), The experience term can significantly improve the model.

```
anova(mod2, mod3, test="Chisq")

## Analysis of Deviance Table
##
## Model 1: call ~ ethnicity + sex
## Model 2: call ~ ethnicity + sex + experience + ethnicity * experienc
e +
##     sex * experience
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4867     2708.9
## 2      4864     2686.4  3   22.573 4.958e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
1 - pchisq(summary(mod2)$deviance - summary(mod3)$deviance, 3)
```

```
## [1] 4.957569e-05
```

6) Construct a plot with *experience* on the horizontal axis and *call* on the vertical axis. Add to this plot four curves, made from the model constructed in question 4. Comment on the similarities or differences between the four curves.
   - For an male applicant with a "Caucasian-sounding" name, add to the plot a red logistic curve showing the probability of getting a call back based on experience.
   - For an female applicant with a "Caucasian-sounding" name, add to the plot a green logistic curve showing the probability of getting a call back based on experience.
   - For a male applicant with an "African-American sounding" name, add to the plot a blue logistic curve showing the probability of getting a call back based on experience.
   - For a female applicant with an "African-American sounding" name, add to the plot a orange logistic curve showing the probability of getting a call back based on experience.

   The curve of both female is trending upward and the curve of both male has a flatting trend, which means that female with more years of experience are more likely to be calling back while to male years of experience has no obvious influence.

   caucasian male applicants have higher calling back rate compared with african american male applicants.

   caucasian female applicants also have higher calling back rate compared with african american female applicants.

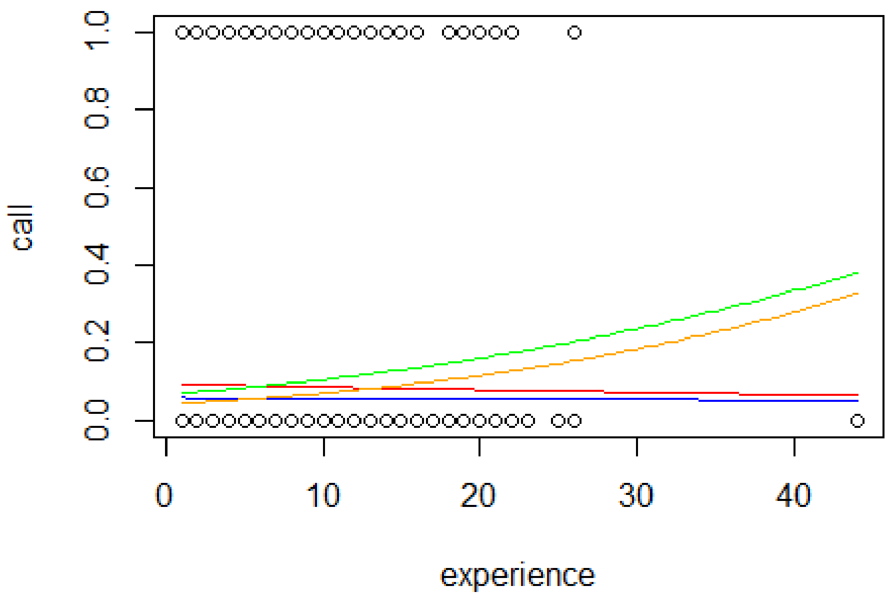
   Overall caucasian applicants have higher calling back rate.

```r
logit = function(B0, B1, x)
  {exp(B0+B1*x)/(1+exp(B0+B1*x))}

plot(call~experience,data=ResumeNames455)

B0 = summary(mod3)$coef[1]
B1 = summary(mod3)$coef[2]
B2 = summary(mod3)$coef[3]
B3 = summary(mod3)$coef[4]
B4 = summary(mod3)$coef[5]
B5 = summary(mod3)$coef[6]
curve(exp(B0+B1+B2+(B3+B4+B5)*x)/(1+exp(B0+B1+B2+(B3+B4+B5)*x)),
                                  add=TRUE, col="red")
curve(exp(B0+B1+(B3+B4)*x)/(1+exp(B0+B1+(B3+B4)*x)),
                                  add=TRUE, col="green")
```

```
curve(exp(B0+B2+(B3+B5)*x)/(1+exp(B0+B2+(B3+B5)*x)),
                                  add=TRUE, col="blue")
curve(exp(B0+B3*x)/(1+exp(B0+B3*x)),
                                  add=TRUE, col="orange")
```



7) Use an appropriate model selection method to construct a best model to predict if the job applicant was called back using any of the variables as predictors (except for *name*). You may also use interaction terms. Why would you not want to use *name* as a predictor?
Different people have different name. There is no significant features or patterns in the words of people's names. The name can only represent that people themselves.

```
library(MASS)
ResumeNames455.1 = within(ResumeNames455,
                     {name = NULL})

fullmod = glm(call~., data=ResumeNames455.1, family="binomial")
final_model_backwards=stepAIC(fullmod, trace=FALSE)
summary(final_model_backwards)

##
## Call:
## glm(formula = call ~ ethnicity + quality + experience, family = "bin
omial",
##     data = ResumeNames455.1)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7103  -0.4340  -0.3939  -0.3428   2.4815
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.915294   0.128516 -22.684  < 2e-16 ***
## ethnicitycauc  0.438628   0.107545   4.079 4.53e-05 ***
## qualitylow    -0.154397   0.106546  -1.449    0.147
## experience     0.037880   0.009256   4.092 4.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2726.9  on 4869  degrees of freedom
## Residual deviance: 2691.1  on 4866  degrees of freedom
## AIC: 2699.1
##
## Number of Fisher Scoring iterations: 5
```