# STOR 455 Homework 6

## 20 points - Due Friday 3/25 5:00pm

### 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

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

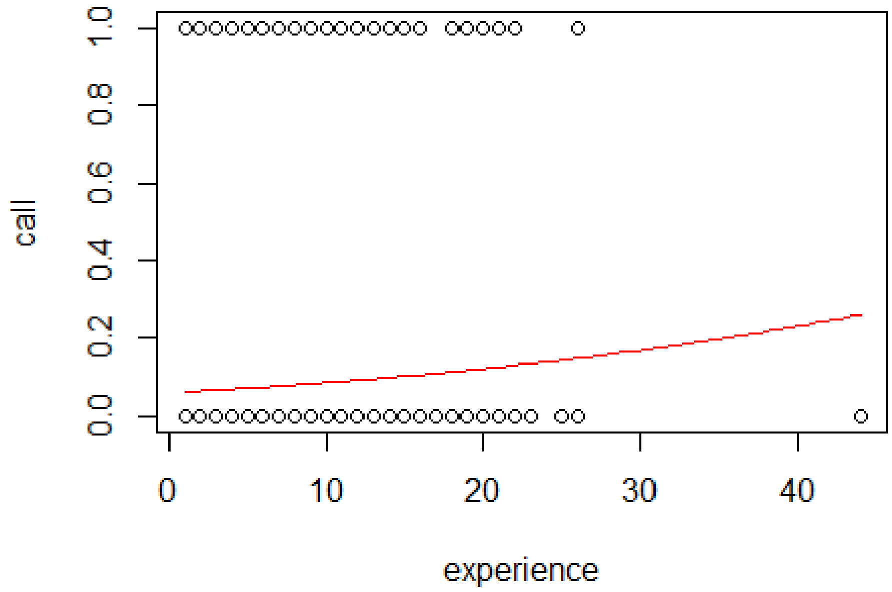
```
## Rows: 4870 Columns: 7

## -- 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) Construct a logistic model to predict if the job applicant was called back using *experience* as the predictor variable.

```
logitmod = glm(call ~ experience, family = binomial, data=ResumeNames45
5)
```

2) Plot the raw data and the logistic curve on the same axes.

```
plot(call ~ experience, data=ResumeNames455)
B0 = summary(logitmod)$coef[1]
B1 = summary(logitmod)$coef[2]
curve(exp(B0+B1*x)/(1+exp(B0+B1*x)),add=TRUE, col="red")
```



3) For an applicant with 6 years of experience, what does your model predict is the probability of this applicant getting called back?

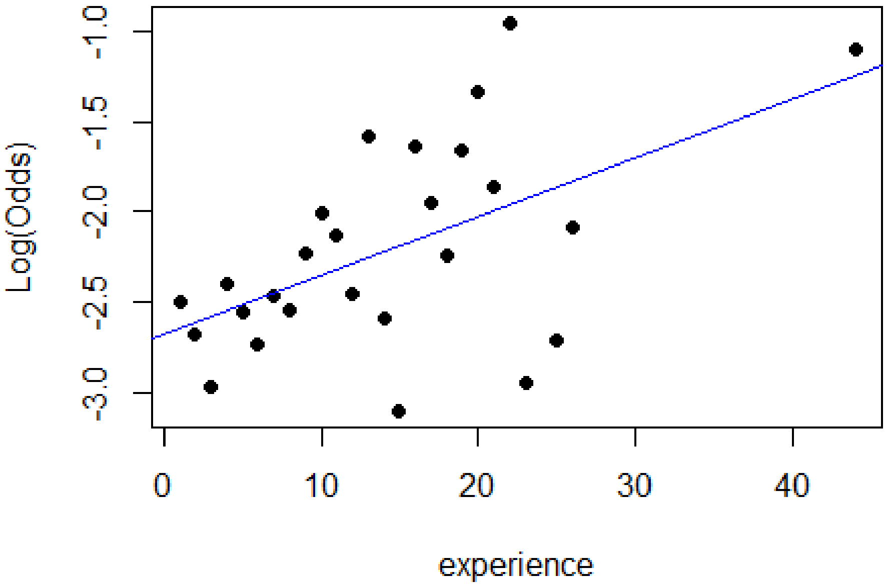
For an applicant with 6 years of experience, my model predict the probability of this applicant getting called back is 0.07411543.

```
newx = data.frame("experience" = 6)
predict(logitmod,newx, type="response")

##          1
## 0.07411543
```

4) Construct an empirical logit plot and comment on the linearity of the data.

The linearity is questionable because the variability from the line follows a fanning pattern as value of the predictor changes.

```
library(Stat2Data)
emplogitplot1(call~experience, data=ResumeNames455, ngroups="all")
```



5) Use the model from question #1 to perform a hypothesis test to determine if there is significant evidence of a relationship between *call* and *experience*. Cite your hypotheses, p-value, and conclusion in context.

H0: Assuming there is no difference in the likelihood of making a plot based on the experience.

Ha: there is some relationship that as the experience of the plot changes, we differently likely to make that plot.

Since P value is pretty small(2.07e-05), we have evidence to say that there is some relationship between call and experience.

```
summary(logitmod)

##
## Call:
## glm(formula = call ~ experience, family = binomial, data = ResumeNam
es455)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7780  -0.4075  -0.3924  -0.3779   2.3598
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.75960    0.09620 -28.687  < 2e-16 ***
## experience   0.03908    0.00918   4.257 2.07e-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: 2710.2  on 4868  degrees of freedom
## AIC: 2714.2
##
## Number of Fisher Scoring iterations: 5
```

6) Construct a confidence interval for the odds ratio for your model and include a sentence interpreting the interval in the context.

We have 95% confidence to say that as the experience increase 1 more year, the odds of being called back increase by a factor of (1.02131169,1.05873170)

```
exp(confint.default(logitmod))

##                  2.5 %     97.5 %
## (Intercept) 0.05243672 0.07645446
## experience  1.02131169 1.05873170
```

7) For each 2-year increase in *experience*, how does your model predict the odds will change for the applicant getting called back?

For each 2-year increase in experience, my model predicts the odds will increase by 1.081295 for the applicant getting called back.

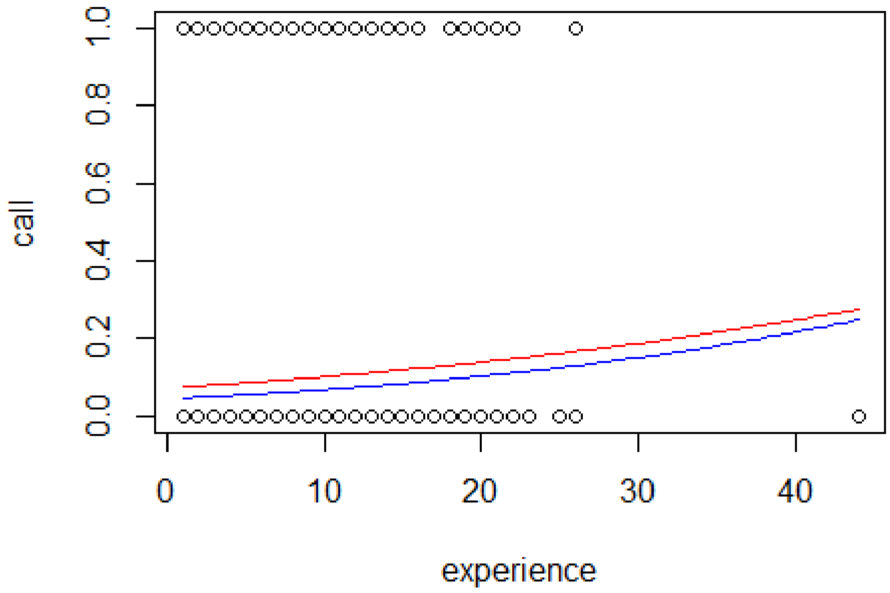
```
B1 = summary(logitmod)$coef[2]
exp(B1*2)

## [1] 1.081295
```

8) Construct subsets of the data for each category of *ethnicity* and construct logistic models to predict if the job applicant was called back using *experience* as the predictor variable for each of these subsets. Then plot the raw data and the logistic curves on the same axes. Comment on differences between the curves and what this means in the context of the data.

From the plot we know that the curve of caucasian applicants is a little bit higher than Afraican-American applicants, which means that under the same condition of years of experience, it is more likely to be called back to a caucasian applicant compared with an Afraican-American applicant.

```
cauc = subset(ResumeNames455, ethnicity=='cauc')
afam = subset(ResumeNames455, ethnicity=='afam')

logitmod_cauc = glm(call~experience, family=binomial, data=cauc)
logitmod_afam = glm(call~experience, family=binomial, data=afam)

plot(call~experience,data=ResumeNames455)

B0_cauc = summary(logitmod_cauc)$coef[1]
```

```
B1_cauc = summary(logitmod_cauc)$coef[2]
curve(exp(B0_cauc+B1_cauc*x)/(1+exp(B0_cauc+B1_cauc*x)),add=TRUE, col="
red")

B0_afam = summary(logitmod_afam)$coef[1]
B1_afam = summary(logitmod_afam)$coef[2]
curve(exp(B0_afam+B1_afam*x)/(1+exp(B0_afam+B1_afam*x)),add=TRUE, col="
blue")
```

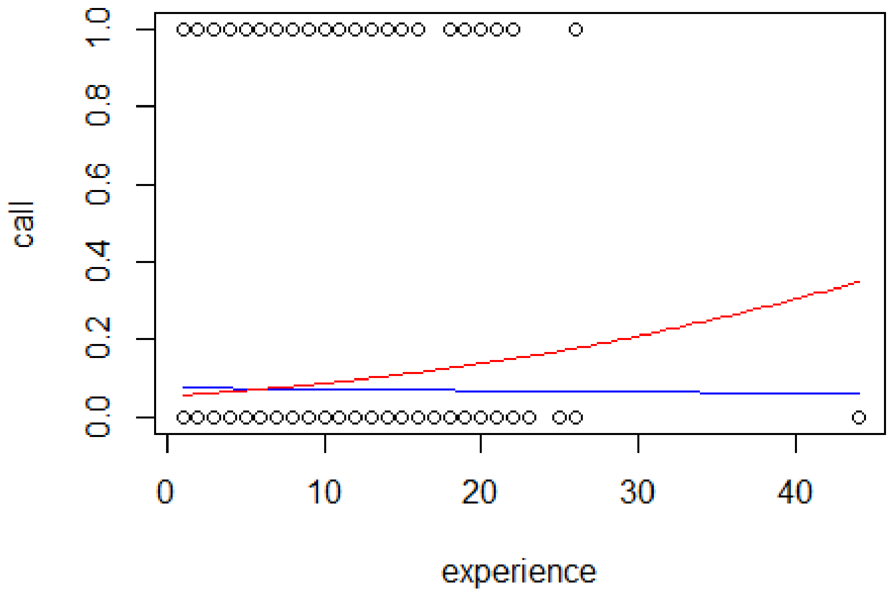


9) Construct subsets of the data for each category of *sex* and construct logistic models to predict if the job applicant was called back using *experience* as the predictor variable for each of these subsets. Then plot the raw data and the logistic curves on the same axes. Comment on differences between the curves and what this means in the context of the data.

From the plot we know that the curve of female is trending upward and the curve of 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.

```
male = subset(ResumeNames455, sex=='male')
female = subset(ResumeNames455, sex=='female')

logitmod_male = glm(call~experience, family=binomial, data=male)
logitmod_female = glm(call~experience, family=binomial, data=female)

plot(call~experience,data=ResumeNames455)
```

```
B0_male = summary(logitmod_male)$coef[1]
B1_male = summary(logitmod_male)$coef[2]
curve(exp(B0_male+B1_male*x)/(1+exp(B0_male+B1_male*x)),add=TRUE, col="
blue")

B0_female = summary(logitmod_female)$coef[1]
B1_female = summary(logitmod_female)$coef[2]
curve(exp(B0_female+B1_female*x)/(1+exp(B0_female+B1_female*x)),add=TRU
E, col="red")
```



In homework #7 we will continue with this data to investigate how the other variables impact an applicant's chances of being called back using multiple logistic regression models.