Final Case Studies

Pei Tao

12/6/2017

Introduction

It's a well known rumor in the job market that there is age discrimination against job seekers; it's said that for most jobs, employers are more keen to hire younger workers as they can work at the company longer as they have more of their career ahead of them, an older and more experience candidate are sometimes considered to be more of an expense as they may request a higher starting salary, and depending on the industry, there is the perception that older workers are actually less capable and less qualified than their younger counterparts due to lack of up-to-date knownledge.

Regardless of if someone believes in the validity of those statements, employment discrimination, where a job-seeker or an employee is treated unfavorably because of his or her race, skin color, national origin, gender, gender identity, disability, religion, sexual orientation or age, is illegal.

Thus when an older female applied for a job as a teacher (she is considered older as she's 40 or older) believes she was not hired because of her age, it is imperative that we find out if the school discriminated her due to her age. The school, on the otherhand, denies the charges, saying that any apparent differences, if present, might be explained by other characteristics of the applicants.

Data Processing

Most of the data cleaning was to clean out the inconsistencies in the words and the "N/A" that litter the dataset. Many of the variables have an extra space after several of the "yes" and "no"s, such as in `interviewed`, `hired`, `GPA`, `MA`, `substitute`, `teaching`, `workkids`, and `residence`, which we replaced, and the "N/A"s were replaced with "NA".

I also seperated the `appdate` into seperate months and years so that it would be useable information in potential linear models.

Since we are very interested in those who are hired, and their age, I also removed the applicants who have value of NA for `interviewed`, `hired`, and `age`.

Finally, I added two columns. One is called `hired_01`, where hired is represented as 1 and not hired is represented as 0. This will allow us to make regressions later on what affects who is `hired`. The other is `interviewed_01`, where 1 represents yes and 0 also represents no.

Analysis

First, let's look at the mean of the ages of those applicants who are hired versus those who are not hired.

```
boxplot(age ~ hired, data=hires,
        xlab = "Hired Status", ylab = "Age", main = "Age of Applicants")
```


```
mean(hires[which(hires$hired == "yes"),]$age)
```

```
## [1] 30.07042
```

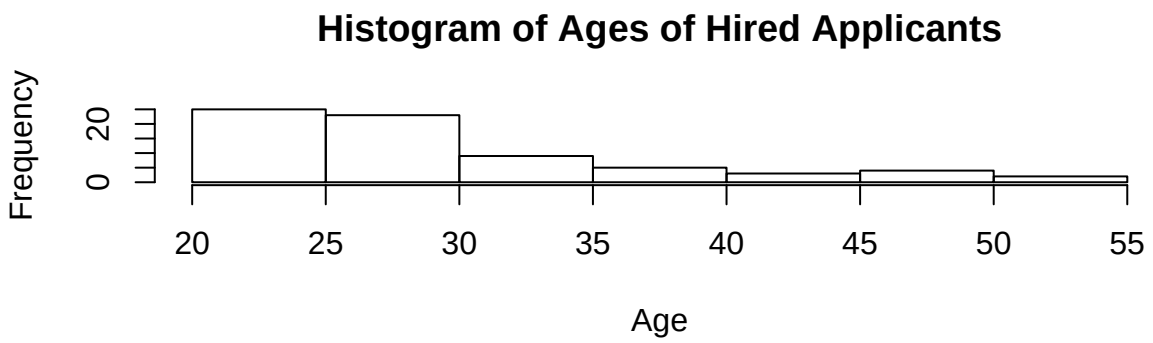
```
mean(hires[which(hires$hired == "no"),]$age)
```

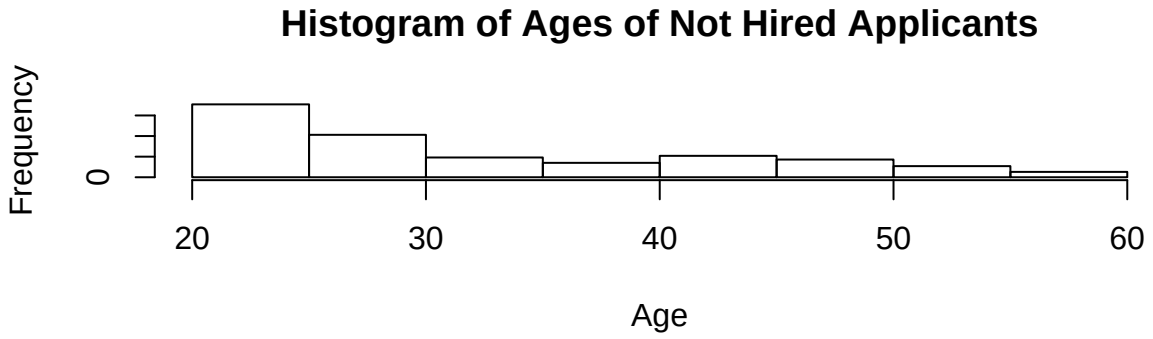
```
## [1] 32.90964
```

The mean age of those who are hired is 30.070424 and the mean age of those who are not hired is 32.90964. This is a difference of almost three years. It seems prudent to test if the different in mean ages is siginificant.

$$H_0 = \text{Mean age of applicants who are hired is the same as those who were not hired}$$

$$H_a = \text{Mean age of the two groups is not the same}$$

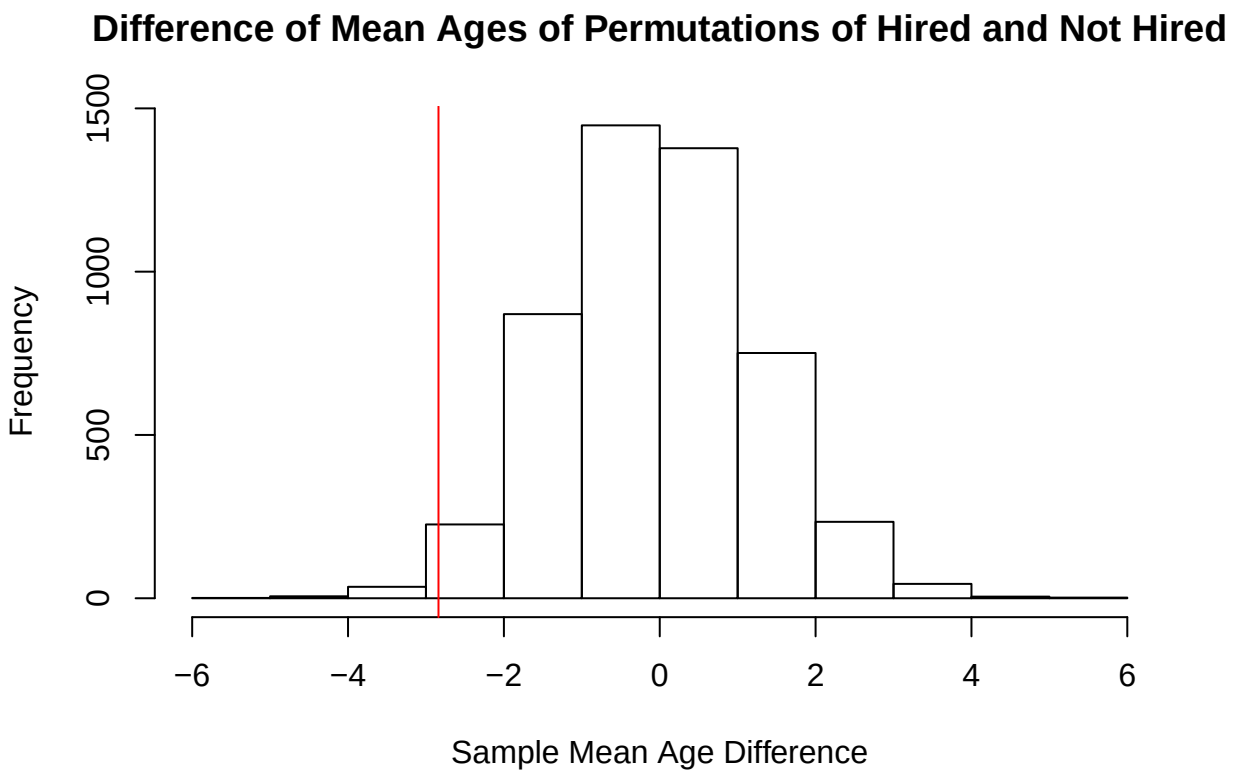




Looking at the histograms of the ages, we see that the ages of the applicants are not normally distributed. Thus instead of doing a t-test, we will perform a permutation test.

```
YEShired_age <- hires[which(hires$hired == "yes"),]$age
NOhired_age <- hires[which(hires$hired == "no"),]$age
test_statistic <- mean(YEShired_age) - mean(NOhired_age)
allages <- c(YEShired_age, NOhired_age)
totlength <- length(YEShired_age) + length(NOhired_age)
N <- 5000
permutate <- rep(0,N)
for (ii in 1:N){
  sample <- sample(allages,totlength,replace=F)
  yes_mean <- mean(sample[1:length(YEShired_age)])
  no_mean <- mean(sample[(length(YEShired_age)+1):totlength])
  permutate[ii] <- yes_mean - no_mean
}
pvalue_permute <- length(which(test_statistic > permutate))/N
```

The p-value we get from this permutation test hovers around 0.01, which is below our alpha level of 0.05. So far, it seems that age may play a factor in the hiring of teachers at this school.

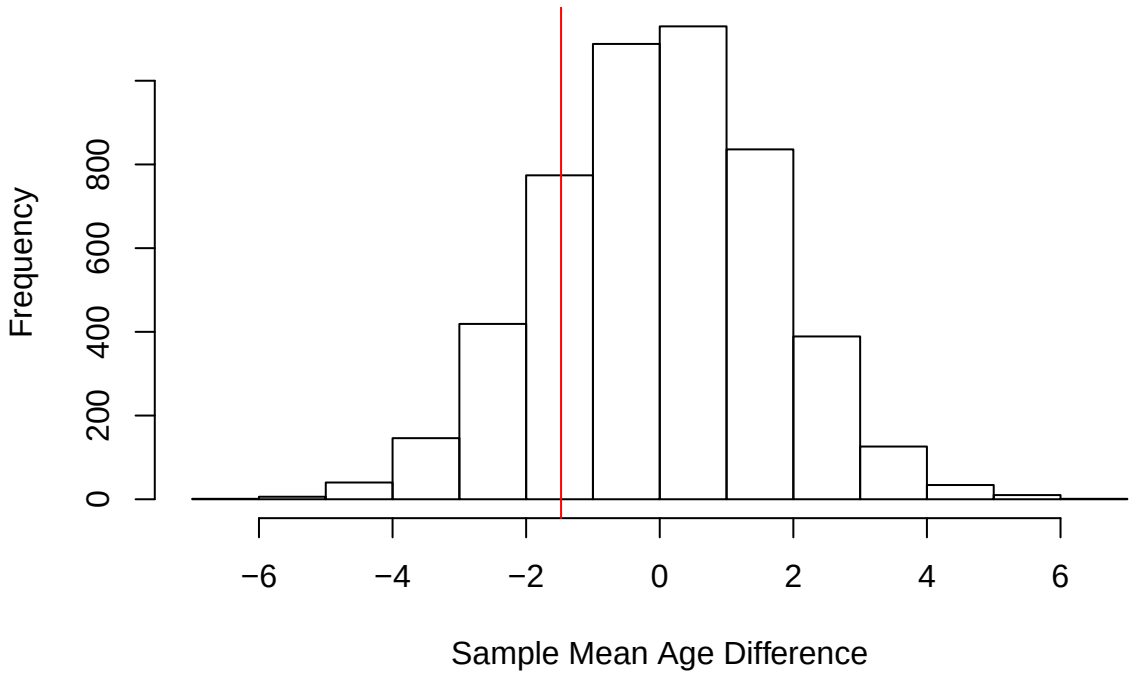


Looking closely at the data, it can be seen that only those who are interviewed have the chance to be hired. My next step was to see that given the applicant is interviewed, is there still an age bias in the hiring process.

H_0 = Mean age of applicants who are interviewed and hired is the same as those who are interviewed but not hired.

H_a = Mean ages are different.

```
interview_hired <- hires[which(hires$interviewed == "yes" & hires$hired == "yes"),]
interview_nothired <- hires[which(hires$interviewed == "yes" & hires$hired == "no"),]
test_statistic_ih <- mean(interview_hired$age) - mean(interview_nothired$age)
allages_ih <- c(interview_hired$age, interview_nothired$age)
totlength_ih <- length(interview_hired$age) + length(interview_nothired$age)
N_ih <- 5000
permutate_ih <- rep(0,N_ih)
for (ii in 1:N_ih){
  sample_ih <- sample(allages_ih,totlength_ih,replace=F)
  yes_mean <- mean(sample_ih[1:length(interview_hired$age)])
  no_mean <- mean(sample_ih[(length(interview_hired$age)+1):totlength_ih])
  permutate_ih[ii] <- yes_mean - no_mean
}
pvalue_ih <- length(which(test_statistic_ih > permutate_ih))/N_ih
```



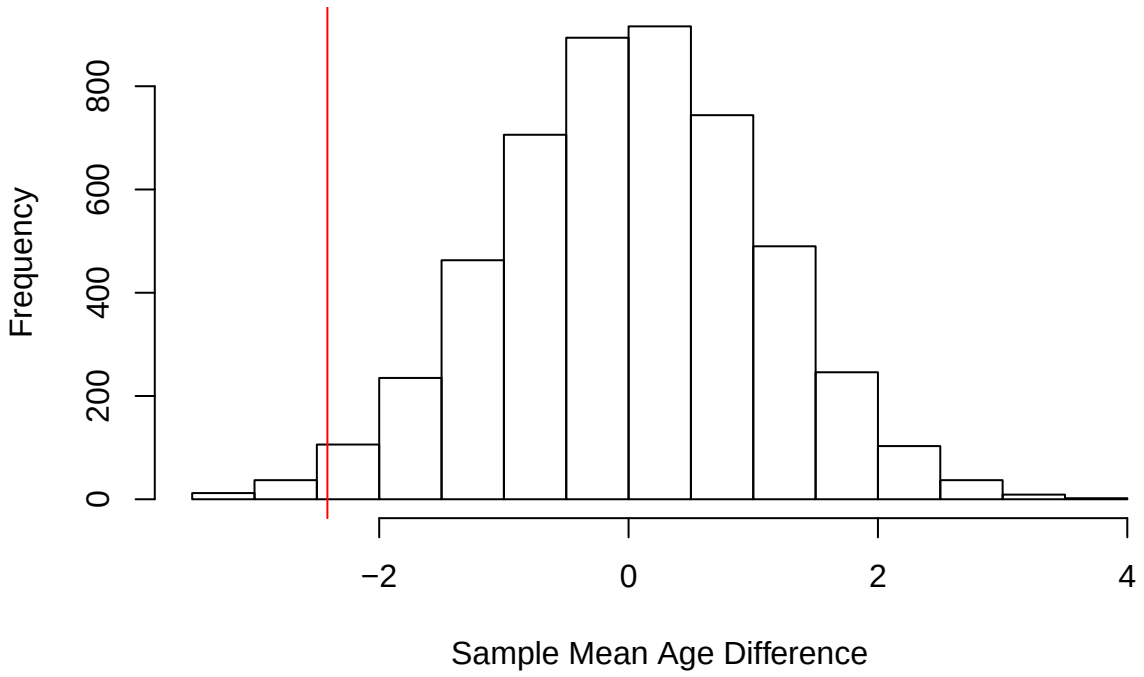
The p-value we get from this permutation test hovers around 0.18, which is above our alpha level of 0.05. Thus given that the candidate is interviewed, it seems like age does not matter in if they get hired. This means that for our original permutation to tell us that age is a significant factor, age must also be significant first in if the applicants gets an interview. Let's check that.

H_0 = Mean age of applicants who are interviewed is the same as those who are not interviewed.

H_a = Mean ages are different.

```
interviewed <- hires[which(hires$interviewed == "yes"),]
not_interviewed <- hires[which(hires$interviewed == "no"),]
test_statistic_interview <- mean(interviewed$age) - mean(not_interviewed$age)
allages_interview <- c(interviewed$age, not_interviewed$age)
totlength_interview <- length(interviewed$age) + length(not_interviewed$age)
N_interview <- 5000
permutate_interview <- rep(0,N_interview)
for (ii in 1:N_interview){
  sample_interview <- sample(allages_interview,totlength_interview,replace=F)
  yes_mean <- mean(sample_interview[1:length(interviewed$age)])
  no_mean <- mean(sample_interview[(length(interviewed$age)+1):totlength_interview])
  permutate_interview[ii] <- yes_mean - no_mean
}
pvalue_interview <- length(which(test_statistic_interview > permutate_interview))/N_interview
```

The p-value of this test is around 0.01 so we reject the null hypothesis in this case. These seems to be an age bias in selecting those to be interviewed.



It would also be interesting to see what other variables are important in hiring a teacher.

From forward stepwise regression thorugh possible variables to use in our logistic regression, we only find the variable `residence` to be a predictor in whether an applicant is hired. backward stepwise regression returns a logistic model of `hired_01 ~ residence + experience + volunteer` and finally, a combination of forward and backward step returns a model of `hired_01 ~ residence`.

```
temp.glm <- glm(formula = hired_01 ~ residence + experience + volunteer, family = binomial,
                data = hires.na)
summary(temp.glm)
```

```
##
## Call:
## glm(formula = hired_01 ~ residence + experience + volunteer,
##     family = binomial, data = hires.na)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0631  -0.4453  -0.3874  -0.3103   2.4730
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.26060    0.44394  -0.587    0.557
## residence2     -1.94825    0.46349  -4.203 2.63e-05 ***
## residence3    -17.01904 1485.98486  -0.011    0.991
## residence4    -17.30547 3956.18035  -0.004    0.997
## experience     -0.05721    0.04001  -1.430    0.153
## volunteeryes    1.22173    0.79132   1.544    0.123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 157.00  on 208  degrees of freedom
## Residual deviance: 134.65  on 203  degrees of freedom
## AIC: 146.65
##
## Number of Fisher Scoring iterations: 16
```

Running the temp.glm reveals that only residence was a significant factor in hiring.

However, regardless of the stepwise regression, I still tried out age as a significant factor in hiring. And it is indeed significant.

```
temp <- glm(hired_01 ~ age,
            data = hires, family=binomial)
summary(temp)
```

```
##
## Call:
## glm(formula = hired_01 ~ age, family = binomial, data = hires)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5925  -0.5756  -0.5194  -0.4151   2.3486
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.95629    0.45523  -2.101   0.0357 *
## age         -0.03157    0.01440  -2.192   0.0284 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 428.28  on 568  degrees of freedom
## Residual deviance: 422.96  on 567  degrees of freedom
## AIC: 426.96
##
## Number of Fisher Scoring iterations: 5
```

However, what we discovered earlier was that age is indeed a significant factor in who is hired, but first, before they are considered for hiring, they are considered for interviews. And age and residence are both significant in who is interviewed.

```
interview.glm <- glm(interviewed_01 ~ age + residence,
                        data = hires, family = binomial)
summary(interview.glm)
```

```
##
## Call:
## glm(formula = interviewed_01 ~ age + residence, family = binomial,
##     data = hires)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.54364  -0.56652  -0.49299  -0.00024   2.32854
##
```

```
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.42528    0.47905   2.975  0.00293 **
## age           -0.02709    0.01330  -2.037  0.04162 *
## residence2    -2.52345    0.26342  -9.579  < 2e-16 ***
## residence3   -18.12885  683.98650  -0.027  0.97885
## residence4   -17.42015 3956.18035  -0.004  0.99649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 567.23  on 568  degrees of freedom
## Residual deviance: 447.41  on 564  degrees of freedom
## AIC: 457.41
##
## Number of Fisher Scoring iterations: 16
```

Finally, once the applicants are interviewed, age is no longer a significant factor, but experience is.

```
interviewed <- hires[which(hires$interviewed == "yes"),]
interviewed.glm <- glm(hired_01 ~ age + residence + experience,
                       data = interviewed, family = binomial)
summary(interviewed.glm)
```

```
##
## Call:
## glm(formula = hired_01 ~ age + residence + experience, family = binomial,
##     data = interviewed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0835  -1.0919   0.5896   0.8856   1.5915
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.14169    1.01417   0.140  0.88889
## age          0.07152    0.04191   1.706  0.08794 .
## residence2  -1.03321    0.43755  -2.361  0.01821 *
## experience  -0.15472    0.05366  -2.883  0.00394 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 143.05  on 109  degrees of freedom
## Residual deviance: 126.77  on 106  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 134.77
##
## Number of Fisher Scoring iterations: 4
```

Conclusion

Both luckily and unluckily for the older female applicant, the data does suggest that age is a significant factor in the hiring process. Although it does not matter in the lawsuit, age is most significant in creating a bias when determining which applicant is to be interviewed. Once they are interviewed, age no longer causes a bias in the hiring process. At that point, experience is a significant factor (though it seems that more experience lowers the chances that the applicant is hired. Plot is provided because it's quite interesting.)

```
boxplot(experience ~ hired, data=hires[which(hires$interviewed == "yes"),],
        main = "Years of Experience of those Interviewed",
        xlab = "Hired",
        ylab = "Years of Experience")
```

The applicant and her lawyer have a case to bring to court.