Case 1: A/B Testing and Experiments, Application to Interviews

Dr. Avery Haviv
University of Rochester
MKT436 and MKT436R

Fall 2020

How to use this case

- Code will be marked using the monospaced Courier New font. For example, we will run regressions with the `lm` function.
- At the end of each section, I will provide some Discussion Questions. In a separate document, I will provide the solutions to these Discussion Questions. To get the most out of the case, I recommend you attempt to solve the questions, in writing, and then check your answer afterwards.
- I will elaborate on some points using footnotes. These footnotes are explicitly not testable material. They might help your understanding or provide some interesting facts.

Introduction

This case centers around an experimental dataset, which was used to study which job applicants were most likely to get a call back. The findings were published in (Bertrand and Mullainathan 2004). The authors generated fake resumes, which they sent to employers who had advertised a position in both Boston and Chicago. They randomized the name on the resume, with some resumes randomly having names that sounded African-American (i.e. Lakisha and Jamal), and the rest sounded White (i.e. Emily and Greg). They also randomized the overall quality of the resumes. **All other attributes of the resume were not fully randomized**.

We will explore this dataset to understand:

1. how to interpret the results of regressions
2. how to analyze experimental data
3. the risks of analyzing non-experimental data

While this study was published in one of the best journal in the world, the analysis itself was relatively simple. This is because when an experiment is properly run, the correct analysis is basic regressions and t-tests.

Load and explore the data

First, we need to load the data. I've posted the data online, and you can download it to your computer. There are several ways to get the dataset in R. The simplest way to do this in RStudio is using the 'import dataset' in the file menu. A preferred, but more advanced method is to set the working directory to a file that includes the data using the `setwd` function, and then using the `read.csv` function to load the data. Given where I stored the data, my code looks like this:

```r
setwd('D:/Dropbox/Teaching Lectures/Interview Case')
resumeData = read.csv('resumeData.csv')
```

You will need to make adapt that code based on where you stored the file. The loaded dataset should have 4870 observations and 27 variables.

We can see the variables in this dataset using the `names` function on the dataset:

```r
names(resumeData)
```

```
##  [1] "name"         "gender"       "ethnicity"    "quality"
##  [5] "call"         "city"         "jobs"         "experience"
##  [9] "honors"       "volunteer"    "military"     "holes"
## [13] "school"       "email"        "computer"     "special"
## [17] "college"      "minimum"      "equal"        "wanted"
## [21] "requirements" "reqexp"       "reqcomm"      "reqeduc"
## [25] "reqcomp"      "reqorg"       "industry"
```

The data guide for this project describes the variables as follows:

- **name** factor indicating applicant's first name.
- **gender** factor indicating gender.
- **ethnicity** factor indicating ethnicity (i.e., Caucasian-sounding vs. African-American sounding first name).
- **quality** factor indicating quality of resume.
- **call** factor. Was the applicant called back?
- **city** factor indicating city: Boston or Chicago.
- **jobs** number of jobs listed on resume.
- **experience** number of years of work experience on the resume.
- **honors** factor. Did the resume mention some honors?
- **volunteer** factor. Did the resume mention some volunteering experience?
- **military** factor. Does the applicant have military experience?
- **holes** factor. Does the resume have some employment holes?
- **school** factor. Does the resume mention some work experience while at school?
- **email** factor. Was the e-mail address on the applicant's resume?
- **computer** factor. Does the resume mention some computer skills?
- **special** factor. Does the resume mention some special skills?
- **college** factor. Does the applicant have a college degree or more?
- **minimum** factor indicating minimum experience requirement of the employer.
- **equal** factor. Is the employer EOE (equal opportunity employment)?
- **wanted** factor indicating type of position wanted by employer.
- **requirements** factor. Does the ad mention some requirement for the job?
- **reqexp** factor. Does the ad mention some experience requirement?
- **reqcomm** factor. Does the ad mention some communication skills requirement?
- **reqeduc** factor. Does the ad mention some educational requirement?

- **reqcomp** factor. Does the ad mention some computer skills requirement?
- **reqorg** factor. Does the ad mention some organizational skills requirement?
- **industry** factor indicating type of employer industry

Using the `summary` function on the dataset gives us more details:

```r
summary(resumeData)
```

```
##       name          gender      ethnicity   quality        call        
##  Tamika : 256   female:3746   afam:2435   high:2446   Mode :logical  
##  Anne   : 242   male  :1124   cauc:2435   low :2424   FALSE:4478     
##  Allison: 232                                         TRUE :392      
##  Latonya: 230                                                        
##  Emily  : 227                                                        
##  Latoya : 226                                                        
##  (Other):3457                                                        
##       city           jobs         experience       honors       
##  boston :2166   Min.   :1.000   Min.   : 1.000   Mode :logical  
##  chicago:2704   1st Qu.:3.000   1st Qu.: 5.000   FALSE:4613     
##                 Median :4.000   Median : 6.000   TRUE :257      
##                 Mean   :3.661   Mean   : 7.843                  
##                 3rd Qu.:4.000   3rd Qu.: 9.000                  
##                 Max.   :7.000   Max.   :44.000                  
##                                                                 
##  volunteer        military         holes           school       
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical  
##  FALSE:2866      FALSE:4397      FALSE:2688      FALSE:2145     
##  TRUE :2004      TRUE :473       TRUE :2182      TRUE :2725     
##                                                                 
##                                                                 
##                                                                 
##                                                                 
##    email          computer        special         college       
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical  
##  FALSE:2536      FALSE:874       FALSE:3269      FALSE:1366     
##  TRUE :2334      TRUE :3996      TRUE :1601      TRUE :3504     
##                                                                 
##                                                                 
##                                                                 
##                                                                 
##     minimum       equal                     wanted     requirements   
##  none   :2746   Mode :logical   manager       : 741   Mode :logical  
##  some   :1064   FALSE:3452      office support: 578   FALSE:1036     
##  2      : 356   TRUE :1418      other         : 736   TRUE :3834     
##  3      : 331                   retail sales  : 818                  
##  5      : 163                   secretary     :1621                  
##  1      : 142                   supervisor    : 376                  
##  (Other):  68                                                        
##    reqexp         reqcomm         reqeduc         reqcomp       
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical  
##  FALSE:2750      FALSE:4262      FALSE:4350      FALSE:2741     
##  TRUE :2120      TRUE :608       TRUE :520       TRUE :2129     
##                                                                 
##                                                                 
```

```
##
##
##    reqorg                                      industry
##  Mode :logical   business/personal services      :1304
##  FALSE:4516      finance/insurance/real estate   : 414
##  TRUE :354       health/education/social services: 754
##                  manufacturing                   : 404
##                  trade                           :1042
##                  transport/communication         : 148
##                  unknown                         : 804
```

Discussion Questions:

1. The data guide describes many variables as being 'factor' variables. What does factor mean in this case? Compare the data guide to the summary to find out
2. Why was `minimum` stored as a factor variable?
3. What is the treatment effect the authors were interested in? What was the outcome variable?

Experimental Variation

As mentioned in the introduction, the name on each resume was randomly assigned to be associated with one of two ethnicities. We are interested in the effect that this variable has on `call`, which indicates whether the resume yielded a call back from the employer. The ethnicity of the resume was stored in the `ethnicity` variable. To see the relationship between these two variables, we can run a simple regression using the `lm` function. The `lm` function requires a formula and a dataset. To write the formula we use the symbol to seperate the left and right hand side of the equation. Since `call` is our dependent variable, it goes to the left of the `" "`, and `ethnicity` is our independent variable, it goes to the right, leading to a formula of `call ethnicity`. Our dataset is still `resumeData`. Notice that we seperate the formula and the data with a comma. We do this anytime there are multiple inputs to a function. Therefore, the regression can be run as follows:

```r
lm(call~ethnicity,data=resumeData)
```

```
##
## Call:
## lm(formula = call ~ ethnicity, data = resumeData)
##
## Coefficients:
##   (Intercept)  ethnicitycauc
##       0.06448        0.03203
```

Our first regression! This exact analysis actually appears in the first row of table 1 of (Bertrand and Mullainathan 2004). Let's interpret the two coefficients. Since `ethnicity` is a categorical variable, it was converted to a binary 0-1 variable, which is 1 if the resume had an Caucasian sounding name.

The intercept is our expected value of call when all other terms are set to 0. In this case, if `ethnicitycauc` is set to 0, the expected value of `call` is 0.06448, or 6.448%. This means that applicants with African-American sounding names got a reply 6.448% of the time.

The coefficient `ethnicitycauc` shows how the expectation changes if `ethnicitycauc` is set to 1. Therefore, the coefficient of 0.03203 means that the probability of getting a call back is increased by 3.203% if the applicant was Caucasian.

Standard Error and Statistical Significance

As you would have learned in statistics class, these coefficients might not be **statistically significant**. A statistically significant coefficient is meaningfully different from 0. We can look at the statistical significance by using the `summary` function on our previous regression:

```
summary(lm(call~ethnicity,data=resumeData))
```

```
##
## Call:
## lm(formula = call ~ ethnicity, data = resumeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09651 -0.09651 -0.06448 -0.06448  0.93552
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.064476   0.005505  11.713  < 2e-16 ***
## ethnicitycauc 0.032033   0.007785   4.115 3.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2716 on 4868 degrees of freedom
## Multiple R-squared:  0.003466,   Adjusted R-squared:  0.003261
## F-statistic: 16.93 on 1 and 4868 DF,  p-value: 3.941e-05
```

Our coefficients are the same as before, but we now have more information. The most important number here is actually the standard error, which tells us how precisely estimated each coefficient is. We can use the standard error to calculate a "confidence interval", which is plausible range of a coefficient. This is calculated to be roughly twice (1.96 times to be exact) the standard error both above and below the estimate. So, in this case, the plausible range for the effect of being Caucasian is $0.032 \pm 1.96 \times 0.07785$, or [0.01677, 0.0473]. Values outside this range are not plausible. In particular, *because 0 is not in the plausible range, this coefficient is statistically significant*. Statistical significance simply tells you how plausible it is that the coefficient is 0.

The coefplot function can help you visualize the range of plausible values (aka the confidence interval) of each coefficient:

```
install.packages('coefplot',repos='http://cran.us.r-project.org')
```

```
## Installing package into 'C:/Users/owner/Documents/R/win-library/3.5'
## (as 'lib' is unspecified)
```

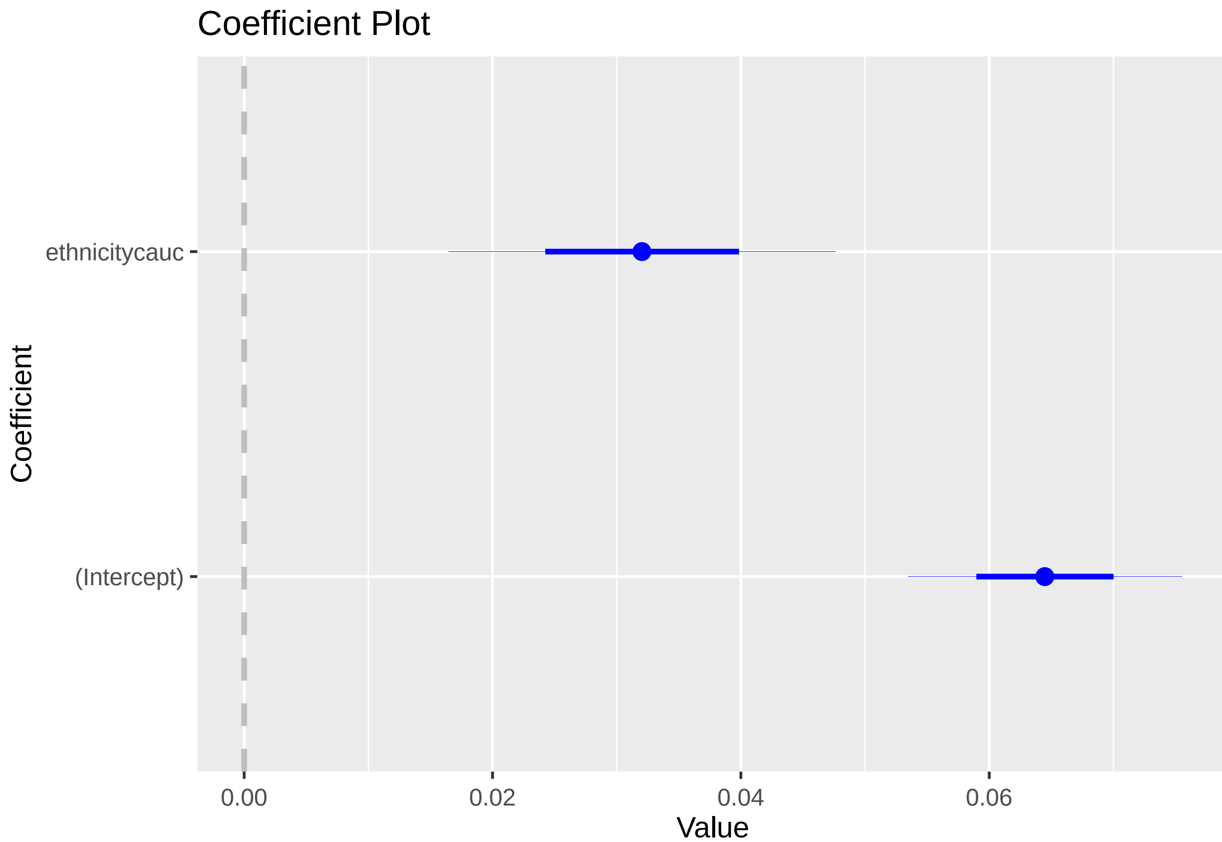
```
## package 'coefplot' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\owner\AppData\Local\Temp\Rtmpgn3xxR\downloaded_packages
```

```
library(coefplot)
```

```
## Warning: package 'coefplot' was built under R version 3.5.3
```

```
## Loading required package: ggplot2
```

```
coefplot(lm(call~ethnicity,data=resumeData))
```



To be clear, even if a coefficient is statistically significant, you still do not know its true value. You can see that a wide range of values for the effect of ethnicity are still plausible.

Benefits of experimental variation

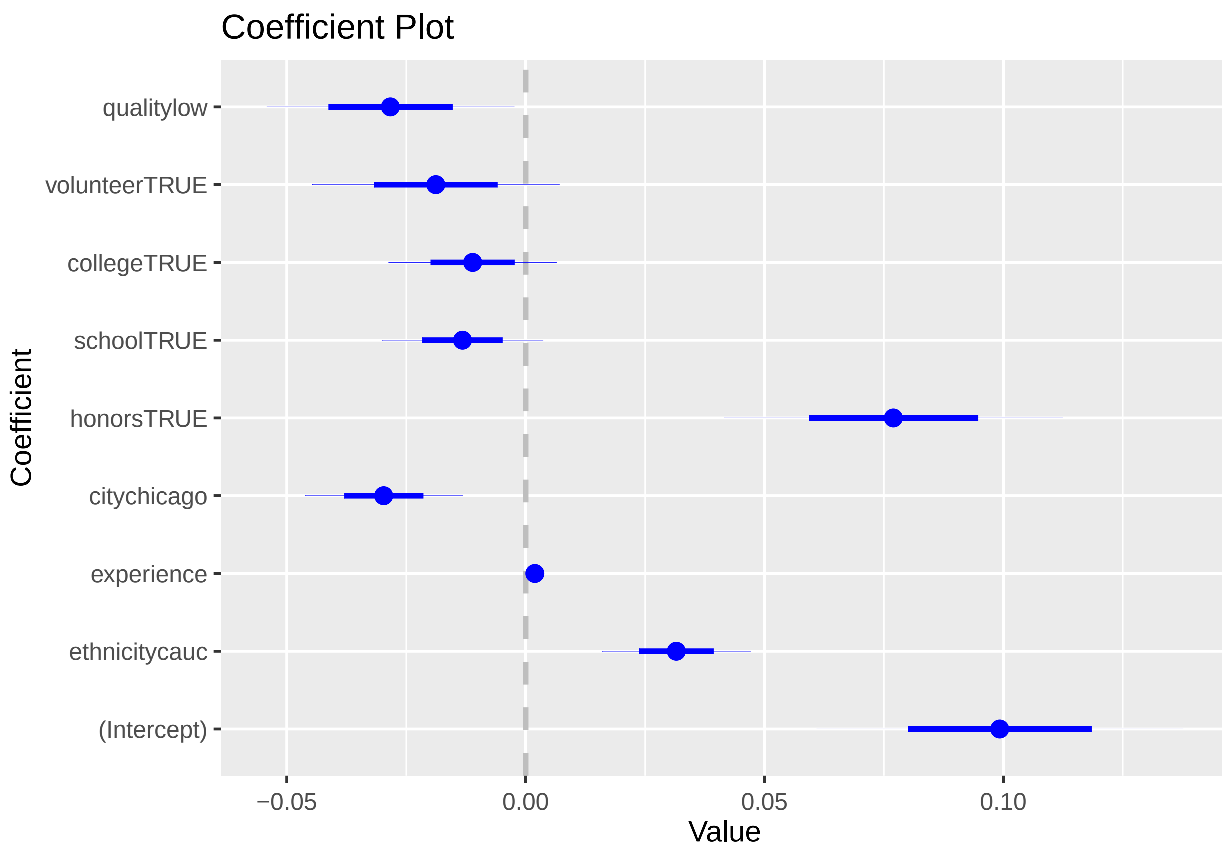
One could try and critique the previous analysis by noting that we did not control for a host of other variables that might affect whether someone gets a call back for an interview. How can we be sure that we know the true effect is due to ethnicity, and not some other variable? For example, employers might prefer candidates with more years of experience, or an honors degree. Below I control for some other variables. Note, I list the other variables I want to control for using the + sign :

```
summary(lm(call~ethnicity+experience+city+honors+school+college+volunteer+quality,data=resumeData))
```

```
##
## Call:
## lm(formula = call ~ ethnicity + experience + city + honors +
##     school + college + volunteer + quality, data = resumeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23910 -0.09359 -0.07307 -0.05120  0.98309
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.0992342  0.0191658   5.178 2.34e-07 ***
```

```
## ethnicitycauc  0.0315411  0.0077460   4.072 4.74e-05 ***
## experience     0.0019288  0.0008057   2.394 0.016709 *
## citychicago   -0.0297144  0.0082257  -3.612 0.000306 ***
## honorsTRUE     0.0769721  0.0176742   4.355 1.36e-05 ***
## schoolTRUE    -0.0132147  0.0084043  -1.572 0.115929
## collegeTRUE   -0.0110911  0.0087987  -1.261 0.207538
## volunteerTRUE -0.0188008  0.0129330  -1.454 0.146093
## qualitylow    -0.0283123  0.0129424  -2.188 0.028749 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2703 on 4861 degrees of freedom
## Multiple R-squared:  0.01503,    Adjusted R-squared:  0.01341
## F-statistic:  9.27 on 8 and 4861 DF,  p-value: 9.191e-13
```

```r
coefplot(lm(call~ethnicity+experience+city+honors+school+college+volunteer+quality,data=resumeData))
```



Clearly many of these variables have significant and important effects. For example, having an honors degree increased the chance of getting a call back by over 8%. Furthermore, the R^2 increased. However, compare the coefficient of `ethnicity` in this regression to the one in the previous section. You will see that the coefficient is largely unchanged. We were able to estimate the effect of ethnicity accurately even *without* all these additional control variables.

This is the biggest benefit of an experiment. When an experiment is properly run, randomization ensures that the only difference, on average, between the treatment and control group is the treatment effect. We can estimate the treatment effect even without controlling for other important factors, and even if the R^2 is low. This is crucially important because in general you *might not have data on the things you need to control for*.

The experiment ensures you don't need to control for anything else.

1. Think about other variables in the dataset that might affect whether someone receives a callback. Test your hypothesis by rerunning the regression above with the additional variables. Did they have the sign you expected? Did adding additional controls affect the coefficient of `ethnicity`?
2. Does `volunteer` have an impact on whether someone receives a callback?
3. According to the estimates in the second analysis, how much would the probability of getting a call back change if a candidate gained two years of experience?

Non-experimental variation

Suppose we wanted to use this dataset to understand the effect that the number of previous jobs had on getting a callback. Similar to the previous section, a simple analysis would run

```
summary(lm(call~jobs,data=resumeData))
```

```
##
## Call:
## lm(formula = call ~ jobs, data = resumeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08221 -0.08067 -0.08015 -0.08015  0.92088
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0786043  0.0123438   6.368 2.09e-10 ***
## jobs        0.0005158  0.0031987   0.161    0.872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2721 on 4868 degrees of freedom
## Multiple R-squared:  5.341e-06,  Adjusted R-squared:  -0.0002001
## F-statistic: 0.026 on 1 and 4868 DF,  p-value: 0.8719
```

From this analysis, it seems that the number of jobs a candidate held has a negligible impact on getting a callback. Note that we can conclude that this impact is small based on both the coefficient and the standard error, as the estimate is small and it is reasonably precise.

What happens if we control for additional variables, as we did previously?

```
summary(lm(call~jobs+experience+city+honors+school+college+volunteer+quality,data=resumeData))
```

```
##
## Call:
## lm(formula = call ~ jobs + experience + city + honors + school +
##     college + volunteer + quality, data = resumeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21823 -0.09269 -0.07408 -0.05356  0.97044
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    0.1430136  0.0235748   6.066 1.41e-09 ***
## jobs          -0.0074244  0.0038756  -1.916   0.0555 .
## experience     0.0023552  0.0008367   2.815   0.0049 **
## citychicago   -0.0364181  0.0089298  -4.078 4.61e-05 ***
## honorsTRUE     0.0707445  0.0180319   3.923 8.86e-05 ***
## schoolTRUE    -0.0116695  0.0084567  -1.380   0.1677
## collegeTRUE   -0.0088255  0.0089114  -0.990   0.3220
## volunteerTRUE -0.0196410  0.0129508  -1.517   0.1294
## qualitylow    -0.0328677  0.0131410  -2.501   0.0124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2706 on 4861 degrees of freedom
## Multiple R-squared:  0.01241,    Adjusted R-squared:  0.01079
## F-statistic: 7.638 on 8 and 4861 DF,  p-value: 3.31e-10
```

Now, the coefficient on jobs is both negative and significant! The reason this can happen is because `jobs` was not randomized: It was correlated with other variables in the dataset. For example, a resume that had more jobs might have more work experience (in years). The initial analysis might have captured the effect of work experience with the jobs variable. Controlling for[1] `experience` removes this possibility. Put differently, without controlling for all relevant variables, we are only getting correlation, not causation.

Now that we've controlled for all these variables, should we be confident that we have the true coefficient of `jobs`? No! Just as there were some variables in our dataset that changed the estimate of `jobs`, there might be variables *not in our dataset* that could similarly change its estimate. Just because you don't have data on a variable, doesn't mean it's not important!

Since `jobs` was not generated from an experiment, the only way to be confident in its coefficient is to carefully think about the things that can be influencing both `jobs` and `call`, and account for them in the analysis. We will be discussing how to do this in later weeks of this class.

Discussion Questions

1. What is the interpretation of the `honorsTRUE` coefficient?
2. Why might an increase in `jobs` lead to a *lower* chance of getting a call back. Isn't having a previous job good? *Hint: You must interpret a coefficient while holding all other variables fixed.*

Randomization Check

The following code checks if there is a significant difference in experience by ethnicty:

```r
summary(lm(experience~ethnicity,data=resumeData))
```

```
##
## Call:
## lm(formula = experience ~ ethnicity, data = resumeData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.856 -2.856 -1.830  1.164 36.170
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.82957    0.10224  76.580   <2e-16 ***
```

[1] AKA including as an independent variable

```
## ethnicitycauc  0.02669    0.14459   0.185    0.854
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.045 on 4868 degrees of freedom
## Multiple R-squared:  7.002e-06,  Adjusted R-squared:  -0.0001984
## F-statistic: 0.03408 on 1 and 4868 DF,  p-value: 0.8535
```

The null hypothesis is that there is no difference in experience. The p-value is 0.8535, which means that there is no statistically significant difference in experience between the two groups. That means it is still plausible that the groups have an equal amount of experience on average, which is what a properly randomized experiment should produce. This analysis is called a 'randomization check', which ensures that the experimental treatment was applied correctly. If the treatment was applied randomly, than it will be uncorrelated with any other potential variable, including those not in the dataset.[2]

We can run a similar test on the `honors` variable:

```r
summary(lm(experience~honors,data=resumeData))
```

```
##
## Call:
## lm(formula = experience ~ honors, data = resumeData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.638 -2.687 -1.687  1.313 36.313
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.68719    0.07364 104.384   <2e-16 ***
## honorsTRUE   2.95094    0.32058   9.205   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.002 on 4868 degrees of freedom
## Multiple R-squared:  0.01711,    Adjusted R-squared:  0.01691
## F-statistic: 84.73 on 1 and 4868 DF,  p-value: < 2.2e-16
```

Here the p-value is very small, and so we can reject the null hypothesis. Applicants with an honors typically have more experience. This means that if we want to estimate the effect of having an honors, we must control for experience. This was a possibility because `honors` was not randomized in this experiment. Note that this is just a demonstration of why we get a consistent estimate of the `ethnicity` coefficient, and we don't get a consistent estimate of other, non-randomized coefficients. You should not use these tests to see if you could control for a variable. If you think a variable might have an effect, simply include it in your analysis.

Discussion Questions

1. Run a regression to see if other variables (i.e. `volunteer` change with `ethnicity`), following the first t-test in this section. Is there a significant difference? Why would you expect this to be the case?
2. Run a t-test to see if other variables (i.e. `volunteer` change with `school`), following the second t-test in this section. Is there a significant difference? Why would you expect this to be the case?

[2] This analysis is actually presented in the second row of table 3 in [@bertrand2004emily]

References

Bertrand, Marianne, and Sendhil Mullainathan. 2004. “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review* 94 (4): 991–1013.