241 Final Project Analysis

Samir Datta

April 14, 2018

1. Load data

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.4.1
```

```
library(ggplot2)
dodge = position_dodge(width=0.9)
theme_set(theme_gray(base_size = 13))

setwd('C:/Users/Samir/Documents/MIDS/ExperimentsCausalityW18')
```

```
ucla_data <- read_excel('data_collection.xlsx', 'User_1_Data_collection')
csulb_data <- read_excel('data_collection.xlsx', 'User_2_Data_collection')
cuny_data <- read_excel('data_collection.xlsx', 'User_3_Data_collection')
ru_data <- read_excel('data_collection.xlsx', 'User_4_Data_collection')

#only include these columns
limited_columns <- c("university_name", "university_brand",
                     "university_state", "good_resume", "job_id",
                             "staggered_application", "date_applied",
                             "call_back", "days_to_respond")

alldata <- rbind(ucla_data[limited_columns],
                 csulb_data[limited_columns],
                 cuny_data[limited_columns],
                 ru_data[limited_columns])

#0 = rejection, 1 = interview, blank/NA = no response
alldata$call_back_factor <- ifelse(is.na(alldata$call_back), "None",
                                   ifelse(alldata$call_back=='0', 'Rejection',
                                          'Interview'))


#binarize call bck variable
alldata$call_back_binary <- ifelse(alldata$call_back_factor=='Interview'&
                                     alldata$days_to_respond<=18,
                                   1,0)

alldata$reject_binary <- ifelse(alldata$call_back_factor=='Rejection'&
                                     alldata$days_to_respond<=18,
                                   1,0)

#states are CA/NY/NJ, condense into west/east coast
alldata$coast <- ifelse(alldata$university_state=='CA', 'West', 'East')
```

```
#count how many jobs had 2 valid applicatons
#deemed valid if the "date_applied" column has just a date or number on t
#and not a blank or comment
alldata$valid <- ifelse(!is.na(as.numeric(alldata$date_applied)), 1, 0)
```

```
## Warning in ifelse(!is.na(as.numeric(alldata$date_applied)), 1, 0): NAs
## introduced by coercion
```

```
valid_agg <- with(alldata, aggregate(valid, list(job_id=job_id), sum))
colnames(valid_agg) <- c('job_id', 'both_applications_valid')
#create binary variable - if a job_id has 2 valid applications, 1, otherwise 0
valid_agg$both_applications_valid <- ifelse(valid_agg$both_applications_valid == 2, 1, 0)

#merge into original data frame
alldata <- merge(alldata, valid_agg, by='job_id')

alldata$phase <- ifelse(alldata$job_id <=80, 'Phase1', 'Phase2')

#get company size from "job openings" sheet
company_info <- read_excel('data_collection.xlsx', 'Job Openings Condensed')

#bin them as S/M/L for 1-49, 50-999, 1000+
company_info$size_bin <- ifelse(is.na(company_info$company_size_n), NA, ifelse(company_info$company_siz
  ifelse(company_info$company_size_n<1000, 'Medium', 'Large')))

#merge into original data frame
alldata <- merge(alldata, company_info[c("job_id", "size_bin", "company_size_n",
                                         "same_city", "same_state", "at_hq_city")])
```

```
print(paste('total # of applicatons sent:', nrow(alldata)))
```

```
## [1] "total # of applicatons sent: 298"
```

```
print(paste('total # of usable applicatons (where both candidates succesfully submitted):', nrow(alldata
```

```
## [1] "total # of usable applicatons (where both candidates succesfully submitted): 210"
```

```
print(paste('total # of usable west coast applicatons:', nrow(alldata[alldata$both_applications_valid==
                                                        alldata$coast=='West',])))
```

```
## [1] "total # of usable west coast applicatons: 96"
```

```
print(paste('total # of usable east coast applicatons:', nrow(alldata[alldata$both_applications_valid==
                                                        alldata$coast=='East',])))
```

```
## [1] "total # of usable east coast applicatons: 114"
```

Covariate checks

```
print(table(alldata[alldata$both_applications_valid==1,]$size_bin,
           alldata[alldata$both_applications_valid==1,]$coast))
```

```
##
##          East West
##   Large    36   46
```

```
##   Medium   54   46
##   Small    24    4
```

```
  chisq.test(alldata[alldata$both_applications_valid==1,]$size_bin,
             alldata[alldata$both_applications_valid==1,]$coast)
```

```
##
##  Pearson's Chi-squared test
##
## data:  alldata[alldata$both_applications_valid == 1, ]$size_bin and alldata[alldata$both_applications
## X-squared = 14.71, df = 2, p-value = 0.0006392
```

We are no longer balanced in terms of the distribution of company size by coast. We found more large west coast companies, and more medium/small east coast companies.

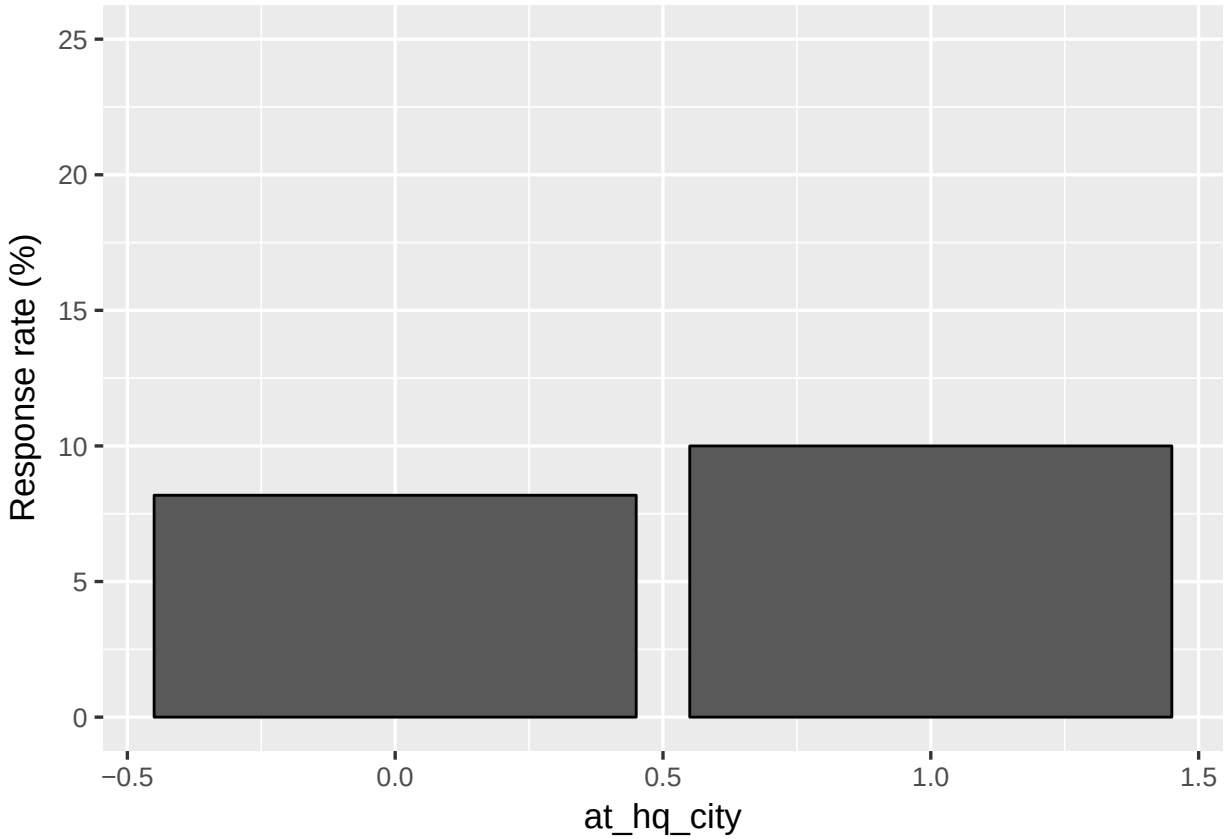
Response rate bar graph - skeleton

```
alldata_agg <- with(alldata[alldata$both_applications_valid==1,],
                    aggregate(job_id,
                              list(call_back_binary=call_back_binary,
                                   at_hq_city=at_hq_city), length))

alldata_agg$response_rate <- NA
alldata_agg[alldata_agg$call_back_binary==1,]$response_rate <-
  100*alldata_agg[alldata_agg$call_back_binary==1,]$x/
  (alldata_agg[alldata_agg$call_back_binary==1,]$x+
     alldata_agg[alldata_agg$call_back_binary==0,]$x)


ggp <- ggplot(alldata_agg[!is.na(alldata_agg$response_rate)], aes(x=at_hq_city, y=response_rate))
ggp + geom_bar(stat="identity", color="black", position=dodge)+
  ylab('Response rate (%)')+ylim(c(0,25))
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

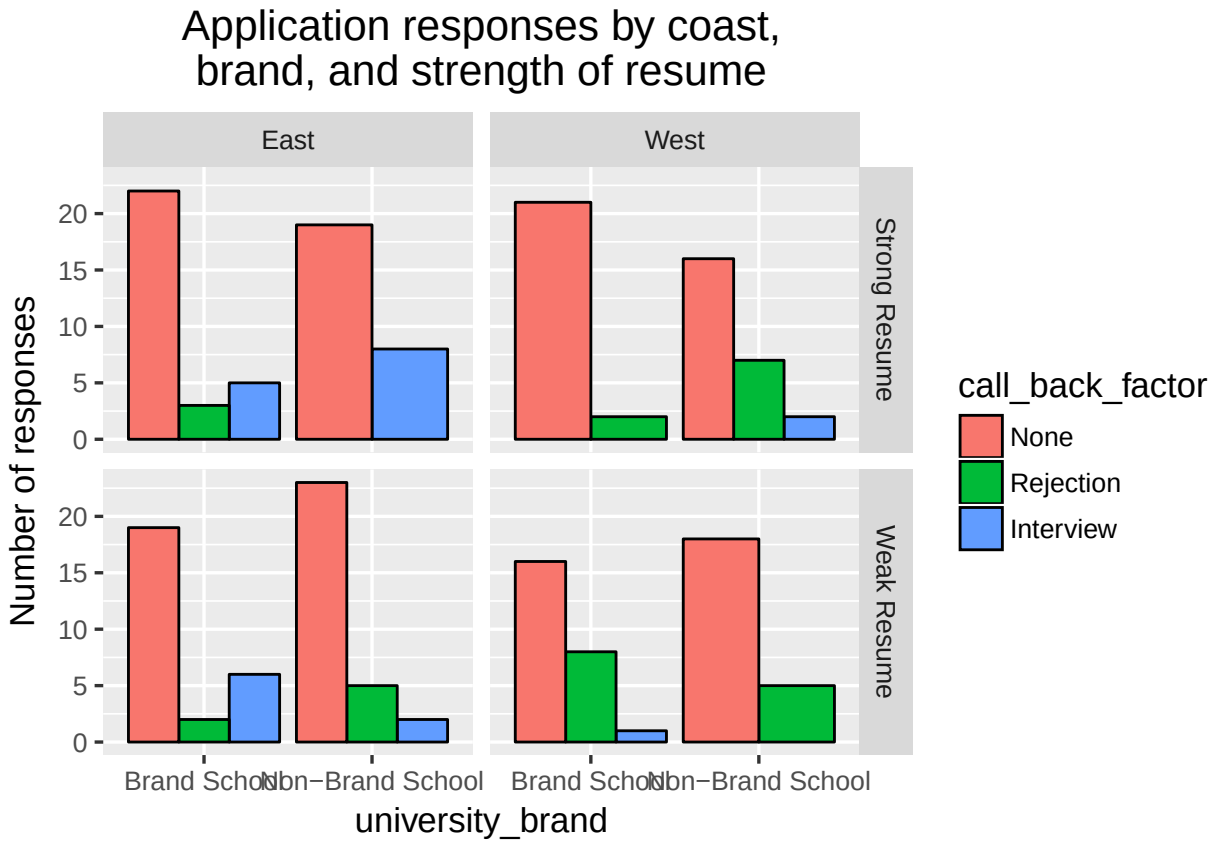



```
  ggtitle('Application response rates')+
  theme_update(plot.title = element_text(hjust = 0.5))
```

```
## NULL
```

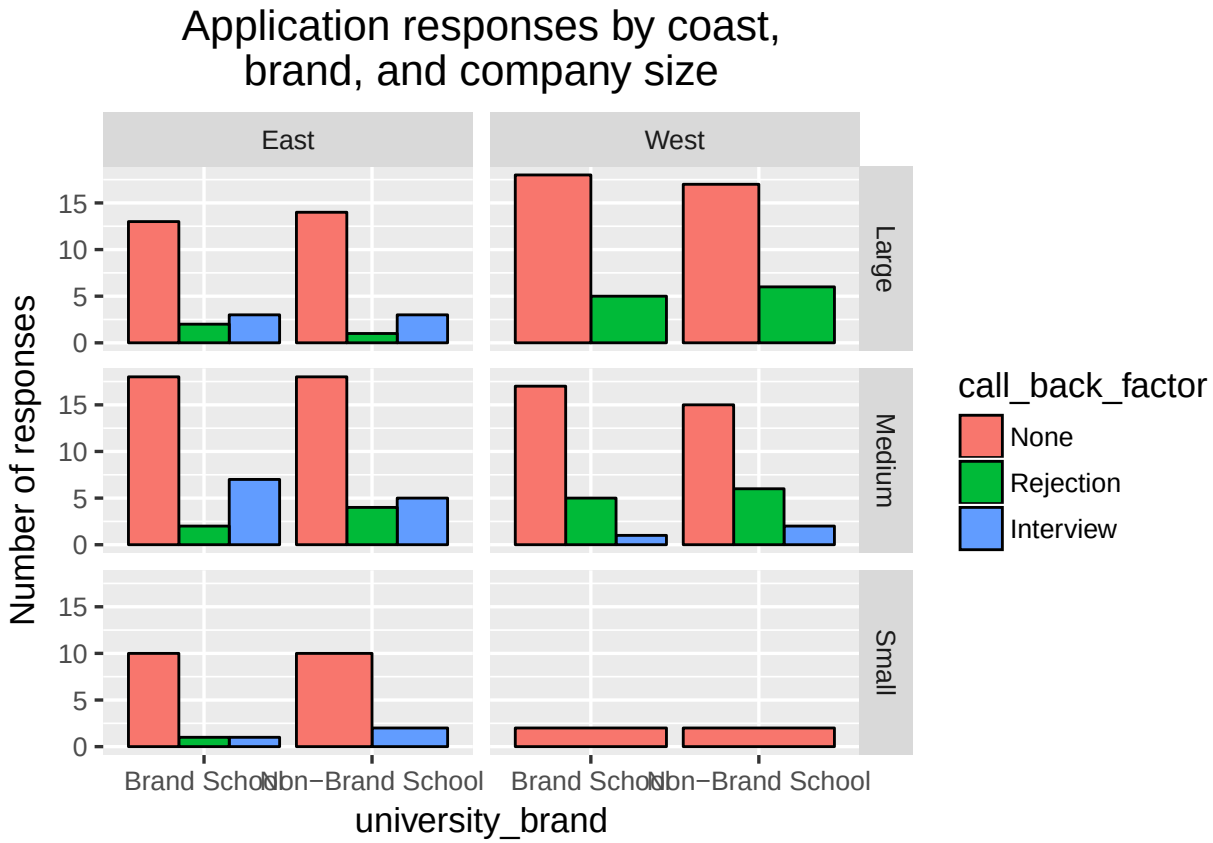
```
alldata_agg <- with(alldata[alldata$both_applications_valid==1,],
                    aggregate(job_id,
                              list(coast=coast, good_resume=good_resume,
                                   call_back_factor=call_back_factor,
                                   university_brand=university_brand), length))
#rename stuff for aesthetics
alldata_agg$call_back_factor <- factor(alldata_agg$call_back_factor,
                                       levels=c('None', 'Rejection', 'Interview'))
alldata_agg$good_resume <- ifelse(alldata_agg$good_resume==1, "Strong Resume",
                                  "Weak Resume")
alldata_agg$university_brand<-ifelse(alldata_agg$university_brand==1, "Brand School",
                                  "Non-Brand School")

ggp <- ggplot(alldata_agg, aes(x=university_brand, y=x,
                               group=call_back_factor, fill=call_back_factor))
ggp + geom_bar(stat="identity", color="black", position=dodge)+
  facet_grid(good_resume~coast)+ylab('Number of responses')+
  ggtitle('Application responses by coast,\nbrand, and strength of resume')+
  theme_update(plot.title = element_text(hjust = 0.5))
```




```
alldata_agg <- with(alldata[alldata$both_applications_valid==1,],
                    aggregate(job_id,
                              list(coast=coast, size_bin=size_bin,
                                   call_back_factor=call_back_factor,
                                   university_brand=university_brand), length))
#rename stuff for aesthetics
alldata_agg$call_back_factor <- factor(alldata_agg$call_back_factor,
                                       levels=c('None', 'Rejection', 'Interview'))
alldata_agg$university_brand<-ifelse(alldata_agg$university_brand==1, "Brand School",
                                     "Non-Brand School")

ggp <- ggplot(alldata_agg, aes(x=university_brand, y=x,
                               group=call_back_factor, fill=call_back_factor))
ggp + geom_bar(stat="identity", color="black", position=dodge)+
  facet_grid(size_bin~coast)+ylab('Number of responses')+
  ggtitle('Application responses by coast,\nbrand, and company size')+
  theme_update(plot.title = element_text(hjust = 0.5))
```



Final model

```
lm.out <- lm(call_back_binary ~ coast+ phase+
               size_bin+staggered_application + university_brand*good_resume, data=alldata[alldata$both_
summary(lm.out)
```

```
##
## Call:
## lm(formula = call_back_binary ~ coast + phase + size_bin + staggered_application +
##     university_brand * good_resume, data = alldata[alldata$both_applications_valid ==
##     1, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27180 -0.13638 -0.07127 -0.00865  0.91547
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.058482   0.060816   0.962   0.3374
## coastWest                    -0.110878   0.040763  -2.720   0.0071 **
## phasePhase2                  -0.033984   0.040493  -0.839   0.4023
## size_binMedium                0.063222   0.043565   1.451   0.1483
## size_binSmall                -0.007071   0.064445  -0.110   0.9127
## staggered_application         0.023146   0.039181   0.591   0.5554
```

```
## university_brand              0.073706   0.056155   1.313   0.1908
## good_resume                   0.126947   0.056204   2.259   0.0250 *
## university_brand:good_resume -0.163355   0.080868  -2.020   0.0447 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2813 on 201 degrees of freedom
## Multiple R-squared:  0.07991,    Adjusted R-squared:  0.04329
## F-statistic: 2.182 on 8 and 201 DF,  p-value: 0.03024
```