Final Report

Experiments and Causality

Colby Carter, Abhishek Agarwal, Tiffany Jaya

April 12, 2018

Load the libraries

```
library(data.table) # fread
library(dplyr)
library(lmtest) # coeftest
library(lubridate) # time conversion
library(sandwich) # vcovHC
```

Helper functions

```
convert_fctr_to_boolean <- function(col) {
  return(as.numeric(as.logical(col)))
}

convert_fctr_to_datetime <- function(col) {
  return(as.POSIXct(col, format="%Y-%m-%d %H:%M:%S"))
}

convert_fctr_to_numeric <- function(col) {
  return(as.numeric(levels(col)[col]))
}

convert_fctr_to_str <- function(col) {
  return(as.character(col))
}
```

Load the dataset

```
# pilot study
d <- read.csv(file = "./W241 Colby Carter, Tiffany Jaya, Abhishek Agarwal_April 5, 2018_00.27.csv",
              header = TRUE,
              sep = ",")
```

```
# actual study
d <- read.csv(file = "./W241 Colby Carter, Tiffany Jaya, Abhishek Agarwal_April 12, 2018_09.41.csv",
              header = TRUE,
              sep = ",")
```

Clean up the dataset

WARNING: DO NOT RERUN THIS SECTION TWICE!

```
# rename columns to be more descriptive
setnames(d,
         old = c("Q1.1",
                 "Q2.2_1",
                 "Q2.3_1",
                 "Q2.5_1",
                 "Q2.7_1",
                 "Q3.1_1",
                 "Q3.3_1",
                 "Q3.5_1",
                 "Q3.7_1",
                 "Q4.1",
                 "Q4.2",
                 "Q4.3",
                 "Q4.4",
                 "Q4.5",
                 "Q4.6",
                 "Q4.7",
                 "Q4.8",
                 "Q4.9",
                 "Q4.10",
                 "Q4.11",
                 "Q4.12",
                 "Q4.13",
                 "Q4.14",
                 "Q4.15",
                 "Q4.16"),
         new = c("consent",
                 "control_employment",
                 "control_education", # baseline
                 "control_retirement",
                 "control_cybersecurity",
                 "treatment_employment",
                 "treatment_education", # baseline,
                 "treatment_retirement",
                 "treatment_cybersecurity",
                 "gender",
                 "age",
                 "highest_education",
                 "employment_status",
                 "marital_status",
                 "zip_code",
                 "community", # rural, urban, suburban
                 "can_vote",
                 "political_party",
                 "ethnicity",
                 "income",
                 "have_kids",
                 "internet_from_mobile",
                 "internet_from_home",
                 "internet_from_work",
                 "who_pays_internet"))
```

```
# 1. remove the first two rows
d <- tail(d, -2)
# 2. remove rownames to avoid confusion (because it is not the subject's id)
rownames(d) <- NULL
# 3. safely convert columns of type factors to their respective types
# 3a. factor -> datetime
cols <- c("StartDate", "EndDate")
d[,cols] <- lapply(d[,cols], convert_fctr_to_datetime)
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'zone/tz/2018c.
## 1.0/zoneinfo/America/Los_Angeles'
```

```
# 3b. factor -> logical/boolean
cols <- c("Finished")
d[,cols] <- lapply(d[,cols], convert_fctr_to_boolean)
```

```
## Warning in `[<-.data.frame`(`*tmp*`, , cols, value = list(1, 1, 1, 1, 1, :
## provided 638 variables to replace 1 variables
```

```
d$consent <- ifelse(d$consent == "Yes", 1, 0)
# 3c. factor -> numeric
cols <- c("Progress", "Duration..in.seconds.")
d[,cols] <- lapply(d[,cols], convert_fctr_to_numeric)
# 3d. factor -> str
cols <- c("IPAddress", "ResponseId", "UserLanguage")
d[,cols] <- lapply(d[,cols], convert_fctr_to_str)
```

```
#d
```

List all the columns

```
colnames(d)
```

```
##  [1] "StartDate"               "EndDate"
##  [3] "Status"                  "IPAddress"
##  [5] "Progress"                "Duration..in.seconds."
##  [7] "Finished"                "RecordedDate"
##  [9] "ResponseId"              "RecipientLastName"
## [11] "RecipientFirstName"      "RecipientEmail"
## [13] "ExternalReference"       "LocationLatitude"
## [15] "LocationLongitude"       "DistributionChannel"
## [17] "UserLanguage"            "consent"
## [19] "Q2.1_First.Click"        "Q2.1_Last.Click"
## [21] "Q2.1_Page.Submit"        "Q2.1_Click.Count"
## [23] "control_employment"      "Q2.4_First.Click"
## [25] "Q2.4_Last.Click"         "Q2.4_Page.Submit"
## [27] "Q2.4_Click.Count"        "control_education"
## [29] "Q2.6_First.Click"        "Q2.6_Last.Click"
## [31] "Q2.6_Page.Submit"        "Q2.6_Click.Count"
## [33] "control_retirement"      "Q2.8_First.Click"
## [35] "Q2.8_Last.Click"         "Q2.8_Page.Submit"
## [37] "Q2.8_Click.Count"        "control_cybersecurity"
## [39] "Q3.2_First.Click"        "Q3.2_Last.Click"
## [41] "Q3.2_Page.Submit"        "Q3.2_Click.Count"
## [43] "treatment_employment"    "Q3.4_First.Click"
## [45] "Q3.4_Last.Click"         "Q3.4_Page.Submit"
## [47] "Q3.4_Click.Count"        "treatment_education"
## [49] "Q3.6_First.Click"        "Q3.6_Last.Click"
## [51] "Q3.6_Page.Submit"        "Q3.6_Click.Count"
## [53] "treatment_retirement"    "Q3.8_First.Click"
## [55] "Q3.8_Last.Click"         "Q3.8_Page.Submit"
## [57] "Q3.8_Click.Count"        "treatment_cybersecurity"
## [59] "gender"                  "age"
## [61] "highest_education"       "employment_status"
## [63] "marital_status"          "zip_code"
## [65] "community"               "can_vote"
## [67] "political_party"         "ethnicity"
## [69] "income"                  "have_kids"
## [71] "internet_from_mobile"    "internet_from_home"
## [73] "internet_from_work"      "who_pays_internet"
## [75] "mTurkCode"               "Q6...Topics"
```

Analysis

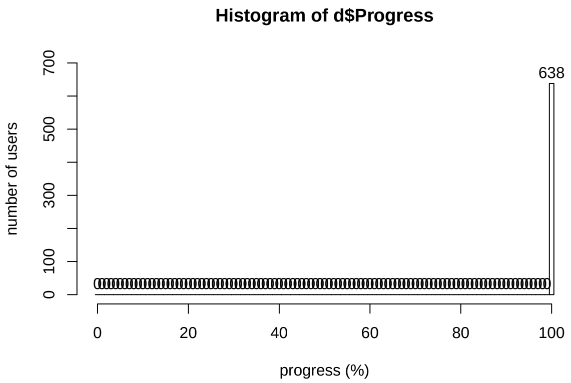
```
# how many subjects take the survey?
nrow(d)
```

```
## [1] 638
```

```
# how many subjects did not finish the survey?
sum(!d$Finished)
```

```
## [1] 0
```

```
hist(d$Progress,
     xlab = "progress (%)",
     ylab = "number of users",
     breaks = -0.5:100.5,
     ylim = c(0, 700),
     labels = TRUE)
```



```
# how long does it take the subjects to finish the survey in minutes?
summary(d$Duration..in.seconds./60)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.050   4.350   5.017   5.539   6.217  20.700
```

```r
# list all subjects who did not consent
subjects.no_consent <- d[which(d$consent == 0),]
nrow(subjects.no_consent)
```

```
## [1] 10
```

```r
# list all the subjects who take less than 2 minutes to finish the survey
subjects.less_2_min <- d[which(d$Duration..in.seconds./60 < 2),]
nrow(subjects.less_2_min)
```

```
## [1] 8
```

```r
# verify that the subjects who did not consent
# match the subjects who took the survey less than 2 minutes
all(subjects.less_2_min$ResponseId %in% subjects.no_consent$ResponseId)
```

```
## [1] TRUE
```

```r
# list all languages not in English
d[which(d$UserLanguage != "EN"),]$UserLanguage
```

```
## [1] ""
```

References

1. renaming column names: http://rprogramming.net/rename-columns-in-r/