Process Pre-Survey, Randomize Treatment, and Check Covariate Balance

Alyssa Eisenberg, Cameron Bell, Sarah Cha

March 18, 2018

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Process Pre-Survey Data

```
# Load the data
data <- read.csv("UCBerkeley TV Habits Study Pre-Survey_March 18, 2018_13.01_CLEAN.csv")

# Remove extraneous columns and rename remaining columns
data <- data[, c(12, 18:36)]
colnames(data) <- c("linkedEmail", "enteredEmail", "gender",
    "age", "region", "employment", "maritalStatus", "children",
    "hoursTV", "binge", "primaryChannel", "allMethods", "moreTimeThanWanted",
    "watchAlone", "shareProfile", "netflixDays", "netflixHours",
    "netflixMin", "netflixAccountAndChrome", "source")

# Generate factors for all multiple choice columns
for (i in c(3:8, 10:11, 13:15, 19)) {
    data[, i] <- factor(data[, i])
}
# Label answers
levels(data$gender) = c("male", "female")
levels(data$age) = c("21-", "22-34", "35-44", "45-54", "55-64",
    "65+")[as.numeric(levels(data$age))]
levels(data$region) = c("midwest", "northeast", "southeast",
    "southwest", "west", "outsideUS")[as.numeric(levels(data$region))]
levels(data$employment) = c("full", "part", "looking", "unemployed",
    "student", "retired", "homemaker", "self", "unable")[as.numeric(levels(data$employment))]
levels(data$maritalStatus) = c("single", "married", "widowed",
    "divorced", "Separated")[as.numeric(levels(data$maritalStatus))]
levels(data$children) = c("yes", "no")[as.numeric(levels(data$children))]
levels(data$binge) = c("once a week", "once a month", "once every couple months",
    "once a year", "no")[as.numeric(levels(data$binge))]
levels(data$primaryChannel) = c("netflix", "HBO", "hulu", "amazon",
    "youtube", "cable", "other")[as.numeric(levels(data$primaryChannel))]
levels(data$moreTimeThanWanted) = c("once a year", "couple times a year",
    "once a month", "couple times a month", "once a week")[as.numeric(levels(data$moreTimeThanWanted))]
levels(data$watchAlone) = c("alone", "withOthers")[as.numeric(levels(data$watchAlone))]
levels(data$shareProfile) = c("yes", "no")[as.numeric(levels(data$shareProfile))]
levels(data$netflixAccountAndChrome) = c("both", "noNetflix",
    "neither", "noChrome")
```

```r
summary(data)
```

```
##                     linkedEmail                       enteredEmail
##  ace9312@gmail.com        : 1   ace9312@gmail.com          : 1
##  alianardo@gmail.com      : 1   alianardo@gmail.com        : 1
##  amodeo@berkeley.edu      : 1   amodeo@berkeley.edu        : 1
##  amycall7@gmail.com       : 1   amycall7@gmail.com         : 1
##  Ariana.viera@gmail.com   : 1   ariana.fonnesbeck@gmail.com: 1
##  bedoukiane@email.chop.edu: 1   bryanmoore@berkeley.edu    : 1
##  (Other)                  :75   (Other)                    :75
##     gender      age           region       employment  maritalStatus
##  male  :36   21-  : 8   midwest  : 4   full     :50   single  :41
##  female:45   22-34:61   northeast:26   part     : 5   married :39
##              35-44: 8   southeast: 4   looking  : 1   divorced: 1
##              45-54: 4   southwest: 2   student  :19
##                         west     :42   homemaker: 3
##                         outsideUS: 3   self     : 3
##
##  children    hoursTV                            binge    primaryChannel
##  yes:17   Min.   : 0.50   once a week              : 7   netflix:54
##  no :64   1st Qu.: 4.00   once a month             :13   HBO    : 1
##           Median : 7.00   once every couple months:28   hulu   : 7
##           Mean   : 8.58   once a year              :12   amazon : 1
##           3rd Qu.:12.00   no                       :21   youtube: 4
##           Max.   :30.00                                  cable  : 4
##                                                          other  :10
##    allMethods            moreTimeThanWanted      watchAlone shareProfile
##  1,4    : 9   once a year         :14       alone     :39   yes:70
##  1,3,4  : 7   couple times a year :27       withOthers:42   no :11
##  1,3,4,5: 7   once a month        :12
##  1,4,5  : 6   couple times a month:17
##  1,2,3,4: 4   once a week         :11
##  1,5    : 4
##  (Other):44
##   netflixDays  netflixHours   netflixMin netflixAccountAndChrome
##  1      :11   Na     :10    Na     :10   both     :57
##  Na     : 9          : 6           : 5   noNetflix:11
##         : 6   7      : 4    36     : 3   neither  : 1
##  0      : 6   0      : 3    21     : 2   noChrome :12
##  2      : 3   21     : 3    6      : 2
##  (Other):10   (Other):20    (Other):25
##  NA's   :36   NA's   :35    NA's   :34
##                    source
##  BYU                  :14
##  Friend               :13
##  Online (social media):26
##  UC Berkeley Slack    :28
##
##
##
```

Randomize treatment

```
# Simple random assignment (treat=1 mean it is in the
# treatment group) set seed so that results of random process
# are reproducible
set.seed(569320)
data$treat <- sample(c(1, 0), size = nrow(data), replace = TRUE)
summary(data$treat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.5185  1.0000  1.0000
```

Covariate balance check

```
# Statistical F-test
model <- lm(treat ~ gender + age + region + employment + maritalStatus +
    children + hoursTV + binge + primaryChannel + moreTimeThanWanted +
    watchAlone + shareProfile + source, data = data)

summary(model)
```

```
##
## Call:
## lm(formula = treat ~ gender + age + region + employment + maritalStatus +
##     children + hoursTV + binge + primaryChannel + moreTimeThanWanted +
##     watchAlone + shareProfile + source, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83086 -0.28271 -0.01469  0.31640  0.90032
##
## Coefficients:
##                                    Estimate Std. Error t value
## (Intercept)                        0.592958   0.634099   0.935
## genderfemale                      -0.015549   0.169939  -0.091
## age22-34                          -0.539260   0.308423  -1.748
## age35-44                          -0.600747   0.424798  -1.414
## age45-54                          -1.306264   0.528002  -2.474
## regionnortheast                    0.267982   0.341704   0.784
## regionsoutheast                   -0.519488   0.471135  -1.103
## regionsouthwest                   -0.295409   0.874745  -0.338
## regionwest                         0.034080   0.349848   0.097
## regionoutsideUS                   -0.151552   0.527816  -0.287
## employmentpart                     0.303626   0.343494   0.884
## employmentlooking                 -0.066978   0.682308  -0.098
## employmentstudent                  0.072856   0.294871   0.247
## employmenthomemaker                0.318297   0.624702   0.510
## employmentself                     0.271009   0.446795   0.607
## maritalStatusmarried               0.158713   0.181386   0.875
## maritalStatusdivorced              0.150948   0.758664   0.199
## childrenno                         0.079323   0.236747   0.335
## hoursTV                            0.003951   0.015881   0.249
```

```
## bingeonce a month                       -0.341357   0.323328  -1.056
## bingeonce every couple months          -0.102676   0.336912  -0.305
## bingeonce a year                        -0.402906   0.391381  -1.029
## bingeno                                  0.074337   0.359537   0.207
## primaryChannelHBO                       -0.595417   0.607137  -0.981
## primaryChannelhulu                      -0.090349   0.276323  -0.327
## primaryChannelamazon                     0.936428   0.621561   1.507
## primaryChannelyoutube                    0.121141   0.385527   0.314
## primaryChannelcable                     -0.066489   0.346772  -0.192
## primaryChannelother                      0.176490   0.232485   0.759
## moreTimeThanWantedcouple times a year  -0.016908   0.246426  -0.069
## moreTimeThanWantedonce a month           0.071240   0.271740   0.262
## moreTimeThanWantedcouple times a month  0.139692   0.254396   0.549
## moreTimeThanWantedonce a week           -0.124344   0.288242  -0.431
## watchAlonewithOthers                    -0.074714   0.173103  -0.432
## shareProfileno                           0.074293   0.302336   0.246
## sourceFriend                             0.241597   0.406697   0.594
## sourceOnline (social media)              0.312077   0.294752   1.059
## sourceUC Berkeley Slack                  0.432445   0.345034   1.253
##                                        Pr(>|t|)
## (Intercept)                              0.3549
## genderfemale                             0.9275
## age22-34                                 0.0875 .
## age35-44                                 0.1645
## age45-54                                 0.0174 *
## regionnortheast                          0.4372
## regionsoutheast                          0.2763
## regionsouthwest                          0.7372
## regionwest                               0.9228
## regionoutsideUS                          0.7754
## employmentpart                           0.3816
## employmentlooking                        0.9223
## employmentstudent                        0.8060
## employmenthomemaker                      0.6130
## employmentself                           0.5473
## maritalStatusmarried                     0.3864
## maritalStatusdivorced                    0.8432
## childrenno                               0.7392
## hoursTV                                  0.8047
## bingeonce a month                        0.2970
## bingeonce every couple months            0.7620
## bingeonce a year                         0.3090
## bingeno                                  0.8372
## primaryChannelHBO                        0.3322
## primaryChannelhulu                       0.7453
## primaryChannelamazon                     0.1392
## primaryChannelyoutube                    0.7549
## primaryChannelcable                      0.8489
## primaryChannelother                      0.4519
## moreTimeThanWantedcouple times a year    0.9456
## moreTimeThanWantedonce a month           0.7944
## moreTimeThanWantedcouple times a month   0.5858
## moreTimeThanWantedonce a week            0.6683
## watchAlonewithOthers                     0.6682
```

```
## shareProfileno                          0.8071
## sourceFriend                            0.5556
## sourceOnline (social media)             0.2956
## sourceUC Berkeley Slack                 0.2169
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5192 on 43 degrees of freedom
## Multiple R-squared:  0.4269, Adjusted R-squared:  -0.06623
## F-statistic: 0.8657 on 37 and 43 DF,  p-value: 0.6711
```

```
# Checked F-stat and p-value. Null hypothesis is that the
# coefficients are jointly equal to 0 We cannot reject the
# null that the variables are jointly insignificant
```

```
# Additionally, examine levels of key covariates by
# treatment/control

se_diff_means <- function(treatment, control) {
    round(sqrt(sd(control)^2/length(control) + sd(treatment)^2/length(treatment)),
        2)
}

# hours of TV
t1 <- round(mean(data$hoursTV[data$treat == 1]), 2)
c1 <- round(mean(data$hoursTV[data$treat == 0]), 2)
diff1 <- t1 - c1
se1 <- se_diff_means(data$hoursTV[data$treat == 1], data$hoursTV[data$treat ==
    0])

# male
t2 <- round(mean(data$gender[data$treat == 1] == "male"), 2)
c2 <- round(mean(data$gender[data$treat == 0] == "male"), 2)
diff2 <- t2 - c2
se2 <- se_diff_means(as.numeric(data$gender[data$treat == 1] ==
    "male"), as.numeric(data$gender[data$treat == 0] == "male"))

# marital status = married
t3 <- round(mean(data$maritalStatus[data$treat == 1] == "married"),
    2)
c3 <- round(mean(data$maritalStatus[data$treat == 0] == "married"),
    2)
diff3 <- t3 - c3
se3 <- se_diff_means(as.numeric(data$maritalStatus[data$treat ==
    1] == "married"), as.numeric(data$maritalStatus[data$treat ==
    0] == "married"))

# no children
t4 <- round(mean(data$children[data$treat == 1] == "no"), 2)
c4 <- round(mean(data$children[data$treat == 0] == "no"), 2)
diff4 <- t4 - c4
se4 <- se_diff_means(as.numeric(data$children[data$treat == 1] ==
    "no"), as.numeric(data$children[data$treat == 0] == "no"))

# moreTimeThanWanted: couple times a year
```

```
t5 <- round(mean(data$moreTimeThanWanted[data$treat == 1] ==
    "couple times a year"), 2)
c5 <- round(mean(data$moreTimeThanWanted[data$treat == 0] ==
    "couple times a year"), 2)
diff5 <- t5 - c5
se5 <- se_diff_means(as.numeric(data$moreTimeThanWanted[data$treat ==
    1] == "couple times a year"), as.numeric(data$moreTimeThanWanted[data$treat ==
    0] == "couple times a year"))

# moreTimeThanWanted: once a month
t6 <- round(mean(data$moreTimeThanWanted[data$treat == 1] ==
    "once a month"), 2)
c6 <- round(mean(data$moreTimeThanWanted[data$treat == 0] ==
    "once a month"), 2)
diff6 <- t6 - c6
se6 <- se_diff_means(as.numeric(data$moreTimeThanWanted[data$treat ==
    1] == "once a month"), as.numeric(data$moreTimeThanWanted[data$treat ==
    0] == "once a month"))

# moreTimeThanWanted: couple times a month
t7 <- round(mean(data$moreTimeThanWanted[data$treat == 1] ==
    "couple times a month"), 2)
c7 <- round(mean(data$moreTimeThanWanted[data$treat == 0] ==
    "couple times a month"), 2)
diff7 <- t7 - c7
se7 <- se_diff_means(as.numeric(data$moreTimeThanWanted[data$treat ==
    1] == "couple times a month"), as.numeric(data$moreTimeThanWanted[data$treat ==
    0] == "couple times a month"))

# moreTimeThanWanted: once a week
t8 <- round(mean(data$moreTimeThanWanted[data$treat == 1] ==
    "once a week"), 2)
c8 <- round(mean(data$moreTimeThanWanted[data$treat == 0] ==
    "once a week"), 2)
diff8 <- t8 - c8
se8 <- se_diff_means(as.numeric(data$moreTimeThanWanted[data$treat ==
    1] == "once a week"), as.numeric(data$moreTimeThanWanted[data$treat ==
    0] == "once a week"))


# Put into a table for display
d <- data.frame(variable = c("hours TV", "male", "married", "no children",
    "watched more than wanted: couple times a year", "watched more than wanted: once a month",
    "watched more than wanted: couple times a month", "watched more than wanted: once a week"),
    control = c(t1, t2, t3, t4, t5, t6, t7, t8), treatment = c(c1,
        c2, c3, c4, c5, c6, c7, c8), diff = c(diff1, diff2, diff3,
        diff4, diff5, diff6, diff7, diff8), se = c(se1, se2,
        se3, se4, se5, se6, se7, se8))
knitr::kable(d)
```

variable	control	treatment	diff	se
hours TV	8.74	8.41	0.33	1.40
male	0.48	0.41	0.07	0.11

variable	control	treatment	diff	se
married	0.43	0.54	-0.11	0.11
no children	0.81	0.77	0.04	0.09
watched more than wanted: couple times a year	0.29	0.38	-0.09	0.11
watched more than wanted: once a month	0.14	0.15	-0.01	0.08
watched more than wanted: couple times a month	0.24	0.18	0.06	0.09
watched more than wanted: once a week	0.14	0.13	0.01	0.08

Finally, output the treatment assignments along with emails.

```
output_data = data[, c("linkedEmail", "treat")]
head(output_data)
```

```
##                 linkedEmail treat
## 1 sarahkelley1759@gmail.com     1
## 2       jennyq.wu@gmail.com     0
## 3    j.fallentine@gmail.com     0
## 4     phat.t.doan@gmail.com     0
## 5    krissy1734@outlook.com     0
## 6       amodeo@berkeley.edu     1
```

```
write.csv(output_data, file = "ExperimentTreatmentAssignment.csv")
```