Final Study Data Analysis

April Kim, Jennifer Podracky, Saurav Datta

```r
library(ggplot2)
library(data.table)
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(pwr)
library(lsr)
```

Read in data and reformat

```r
d2 <- read.csv("241 Participant List - Final Study Results - 20181215.csv",
                     stringsAsFactors = F, na.strings=c("","NA"))

d2$Q7 <- as.numeric(gsub("\\,", "", d2$Q7))
```

```
## Warning: NAs introduced by coercion
```

```r
d2$Q11 <- as.numeric(gsub("\\,", "", d2$Q11))
```

```
## Warning: NAs introduced by coercion
```

```r
d2$Q15 <- as.numeric(gsub("\\,", "", d2$Q15))
```

```
## Warning: NAs introduced by coercion
```

```r
# Not applicable = 0
# Through digital means = 1
# In person = 2
# Both in person and through digital means = 3

d2[d2$Q6 %like% "Not applicable", ]$Q6 <- 0
d2[d2$Q10 %like% "Not applicable", ]$Q10 <- 0
d2[d2$Q14 %like% "Not applicable", ]$Q14 <- 0

d2[d2$Q6 %like% "In person", ]$Q6 <- 2
d2[d2$Q10 %like% "In person", ]$Q10 <- 2
```

```r
d2[d2$Q14 %like% "In person", ]$Q14 <- 2

d2[d2$Q6 %like% "Through digital means", ]$Q6 <- 1
d2[d2$Q10 %like% "Through digital means", ]$Q10 <- 1
d2[d2$Q14 %like% "Through digital means", ]$Q14 <- 1

d2[d2$Q6 %like% "Both in person and through digital means", ]$Q6 <- 3
d2[d2$Q10 %like% "Both in person and through digital means", ]$Q10 <- 3
# d2[d2$Q14 %like% "Both in person and through digital means", ]$Q14 <- 3

d2$Q6 <- as.numeric(as.character(d2$Q6))
d2$Q10 <- as.numeric(as.character(d2$Q10))
d2$Q14 <- as.numeric(as.character(d2$Q14))

names(d2) <- c("user_id", "treatment_seq", "location_lat", "location_long", "name", "email",
               "gender", "age_range", "location", "lives_with_others", "submitted_response", "noncompli
               "day1_num_persons", "day1_treatment", "day1_steps", "day1_upload_id", "day1_upload_filen
               "day2_num_persons", "day2_treatment", "day2_steps", "day2_upload_id", "day2_upload_filen
               "day3_num_persons", "day3_treatment", "day3_steps", "day3_upload_id", "day3_upload_filen
```

Unless I completely misunbderstood Alex's point during OH, if we check for ordering effect in the way he suggested, there's going to be exact collinearity and coefficient is not estimable. Jenni suggested testing if previous day's treatment is highly predictive of next day's step count (thank you Jenni!)

Checking for ordering/priming effect Is previous day's treatment highly predictive of how many steps are taken today?

```r
# downsize to fewer columns so as to not update the original data
# and it's easier look through :)
# n = 75
df <- d2[,c(2,5,7,8,15,16,22,23,29,30)]

# remove subjects/rows who were noncompliant
# n = 24
df <- df[rowSums(is.na(df[,c(5:10)])) != ncol(df[,c(5:10)]), ]

# day 3 steps using day 1 and 2 treatment
m1 <- lm(day3_steps ~ day1_treatment + day2_treatment, df)
summary(m1)
```

```
##
## Call:
## lm(formula = day3_steps ~ day1_treatment + day2_treatment, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7212.3 -2119.9  -622.5  1181.3  8810.7
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7224.4      936.6   7.714 9.03e-10 ***
```

```
## day1_treatment    -12.2      620.8  -0.020    0.984
## day2_treatment   -340.1      606.5  -0.561    0.578
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3452 on 45 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.007005,   Adjusted R-squared:  -0.03713
## F-statistic: 0.1587 on 2 and 45 DF,  p-value: 0.8537
```

```r
# ATE (standard error)
print(paste0("Estimated effect of day1 treatment: ", signif(m1$coefficients[2], 3),
             " (", signif(coef(summary(m1))[2,2], 3), ")"))
```

```
## [1] "Estimated effect of day1 treatment: -12.2 (621)"
```

```r
print(paste0("Estimated effect of day2 treatment: ", signif(m1$coefficients[3], 3),
             " (", signif(coef(summary(m1))[3,2], 3), ")"))
```

```
## [1] "Estimated effect of day2 treatment: -340 (607)"
```

```r
# include days1,2 steps as covariates to uhderstand
# subjects' step counts hange as a function of
# treatment against waht they would typically do
m2 <- lm(day3_steps ~ day1_treatment + day2_treatment + day1_steps + day2_steps, df)
summary(m2)
```

```
##
## Call:
## lm(formula = day3_steps ~ day1_treatment + day2_treatment + day1_steps +
##     day2_steps, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8250.2 -1347.6    -5.5  1539.5  5017.0
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2457.4095  1124.7333   2.185  0.03453 *
## day1_treatment -297.2370   482.7203  -0.616  0.54138
## day2_treatment -304.0639   467.0271  -0.651  0.51855
## day1_steps        0.2382     0.1241   1.920  0.06166 .
## day2_steps        0.4649     0.1361   3.417  0.00142 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2639 on 42 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.4486, Adjusted R-squared:  0.3961
## F-statistic: 8.542 on 4 and 42 DF,  p-value: 3.881e-05
```

```r
print(paste0("Estimated effect of day1 treatment: ", signif(m2$coefficients[2], 3),
             " (", signif(coef(summary(m2))[2,2], 3), ")"))
```

```
## [1] "Estimated effect of day1 treatment: -297 (483)"
```

```r
print(paste0("Estimated effect of day2 treatment: ", signif(m2$coefficients[3], 3),
             " (", signif(coef(summary(m2))[3,2], 3), ")"))
```

```
## [1] "Estimated effect of day2 treatment: -304 (467)"
```

We do not see that the previous days' treatment assignments to predict the last day's step count is highgly predicitive and significant, which is super for us!

Condense treatment sequence to 1 treatment

```r
# get data frame in workable format
d <- rbindlist(list(df[1:nrow(df),c("day1_treatment","day1_steps")],
                    df[1:nrow(df),c("day2_treatment","day2_steps")],
                    df[1:nrow(df),c("day3_treatment","day3_steps")]))
names(d) <- c("treatment","steps")
d$outcome <- ifelse(d$steps > 5000, 1, 0)

ITT_lm <- lm(outcome ~ treatment, data = d)
summary(ITT_lm)$coefficients[2]
```

```
## [1] 0.01912468
```

```r
# combine digital and in person as one treatment as
# "Both in person and through digital means"
d1 <- d
d1[d1$treatment == 1 | d1$treatment == 2, ]$treatment <- 3
```

T-test and power calculations

```r
###  Control vs digital
t.test(d[treatment == 0]$outcome, d[treatment == 1]$outcome, paired = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  d[treatment == 0]$outcome and d[treatment == 1]$outcome
## t = -1.4806, df = 84.558, p-value = 0.1424
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.29436977  0.04308772
## sample estimates:
## mean of x mean of y
## 0.7076923 0.8333333
```

```r
cohensD(d[treatment == 0]$outcome, d[treatment == 1]$outcome, method = "unequal")
```

```
## [1] 0.2990803
```

```r
### Control vs in person
t.test(d[treatment == 0]$outcome, d[treatment == 2]$outcome, paired = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  d[treatment == 0]$outcome and d[treatment == 2]$outcome
## t = -0.072768, df = 87.814, p-value = 0.9422
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1866635  0.1734767
## sample estimates:
## mean of x mean of y
## 0.7076923 0.7142857
```

```r
cohensD(d[treatment == 0]$outcome, d[treatment == 2]$outcome, method = "unequal")
```

```
## [1] 0.01440248
```

```r
### Control vs in person + digital
t.test(d1[treatment == 0]$outcome, d1[treatment == 3]$outcome, paired = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  d1[treatment == 0]$outcome and d1[treatment == 3]$outcome
## t = -0.91261, df = 131.56, p-value = 0.3631
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.21320328  0.07858789
## sample estimates:
## mean of x mean of y
## 0.7076923 0.7750000
```

```r
cohensD(d1[treatment == 0]$outcome, d1[treatment == 3]$outcome, method = "unequal")
```

```
## [1] 0.1530752
```

```r
### In person vs digital
t.test(d[treatment == 2]$outcome, d[treatment == 1]$outcome, paired = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  d[treatment == 2]$outcome and d[treatment == 1]$outcome
```

```
## t = -1.2587, df = 75.912, p-value = 0.212
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.30742857  0.06933333
## sample estimates:
## mean of x mean of y
## 0.7142857 0.8333333
```

no significant difference in both comparisons