Multiple Regression

Matthew J. Salganik

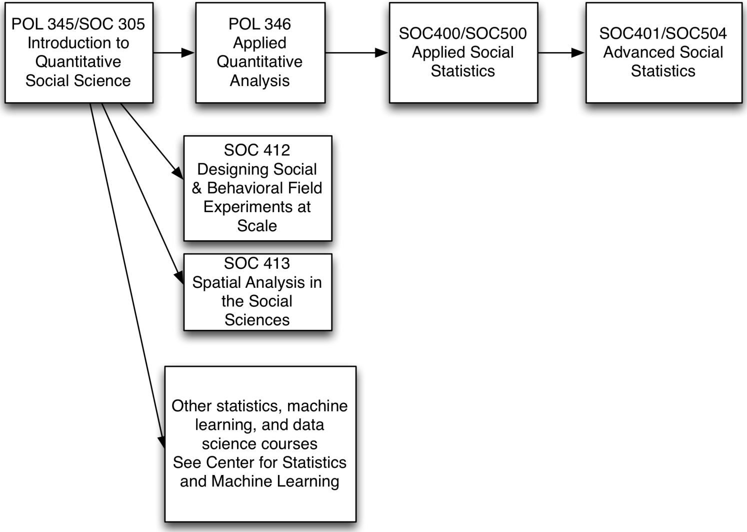
POL 345/SOC 305
Introduction to Quantitative Social Science
Princeton University

Wednesday, November 29, 2017

Your future courses

Omar Wasow from POL 346 Applied Quantitative Analysis

Your future courses



POL 345/SOC 305 Introduction to Quantitative Social Science

POL 346 Applied Quantitative Analysis

SOC400/SOC500 Applied Social Statistics

SOC401/SOC504 Advanced Social Statistics

SOC 412 Designing Social & Behavioral Field Experiments at Scale

SOC 413 Spatial Analysis in the Social Sciences

Other statistics, machine learning, and data science courses
See Center for Statistics and Machine Learning

Logistics

- QSS assignments due 24 hours before precept

Logistics

- QSS assignments due 24 hours before precept
- Pset 3 will be posted W 12/6 and due W 12/13

Logistics

- QSS assignments due 24 hours before precept
- Pset 3 will be posted W 12/6 and due W 12/13
- COMPASS workshop: Thurs, 11/30 Text Mining in R (Ethan)

Goals for today

- See real data analyis workflow (with data wrangling)

Goals for today

- See real data analyis workflow (with data wrangling)
- Review difference-of-means

Goals for today

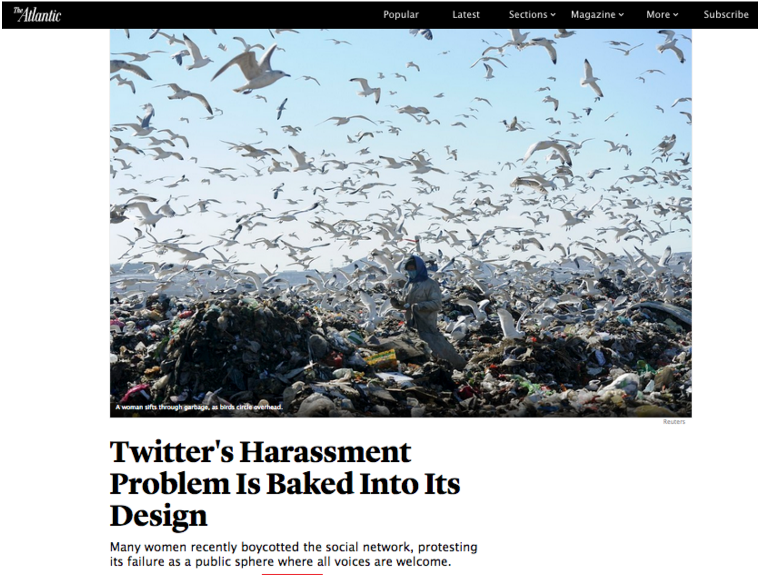
- See real data analyis workflow (with data wrangling)
- Review difference-of-means
- Show connection between difference-of-means and regression

Goals for today

- See real data analyis workflow (with data wrangling)
- Review difference-of-means
- Show connection between difference-of-means and regression
- Explore multiple regression with continuous and dummy variables in equations, code, pictures, and words

Goals for today

- See real data analyis workflow (with data wrangling)
- Review difference-of-means
- Show connection between difference-of-means and regression
- Explore multiple regression with continuous and dummy variables in equations, code, pictures, and words
- Learn something about Twitter



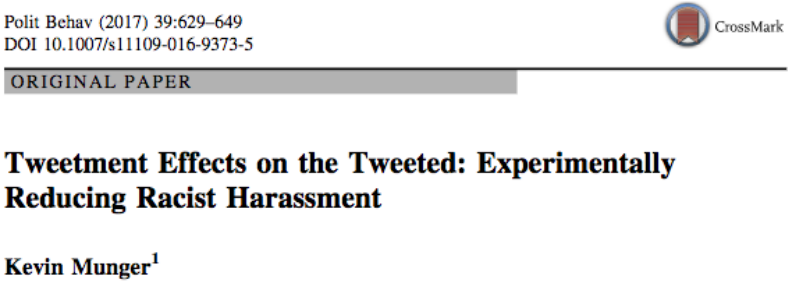
A woman sifts through garbage, as birds circle overhead.

Reuters

Twitter's Harassment Problem Is Baked Into Its Design

Many women recently boycotted the social network, protesting its failure as a public sphere where all voices are welcome.

https://www.theatlantic.com/technology/archive/2017/10/
twitters-harassment-problem-is-baked-into-its-design/542952/



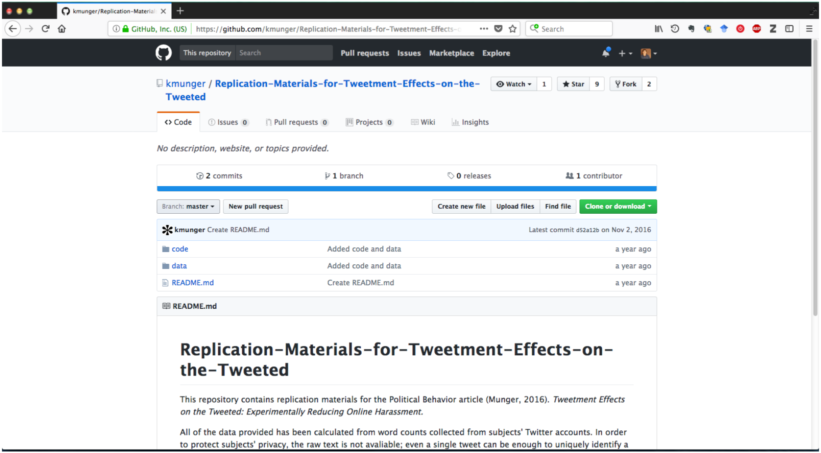
Polit Behav (2017) 39:629–649
DOI 10.1007/s11109-016-9373-5

ORIGINAL PAPER

Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

Kevin Munger[1]

http://dx.doi.org/10.1007/s11109-016-9373-5



kmunger / Replication-Materials-for-Tweetment-Effects-on-the-Tweeted

No description, website, or topics provided.

2 commits · 1 branch · 0 releases · 1 contributor

kmunger Create README.md — Latest commit d52a12b on Nov 2, 2016

code	Added code and data	a year ago
data	Added code and data	a year ago
README.md	Create README.md	a year ago

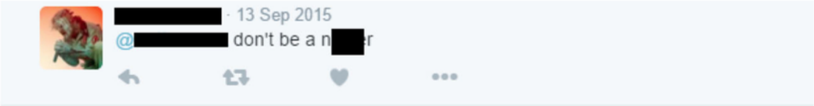
README.md

Replication-Materials-for-Tweetment-Effects-on-the-Tweeted

This repository contains replication materials for the Political Behavior article (Munger, 2016). *Tweetment Effects on the Tweeted: Experimentally Reducing Online Harassment.*

All of the data provided has been calculated from word counts collected from subjects' Twitter accounts. In order to protect subjects' privacy, the raw text is not avaliable; even a single tweet can be enough to uniquely identify a

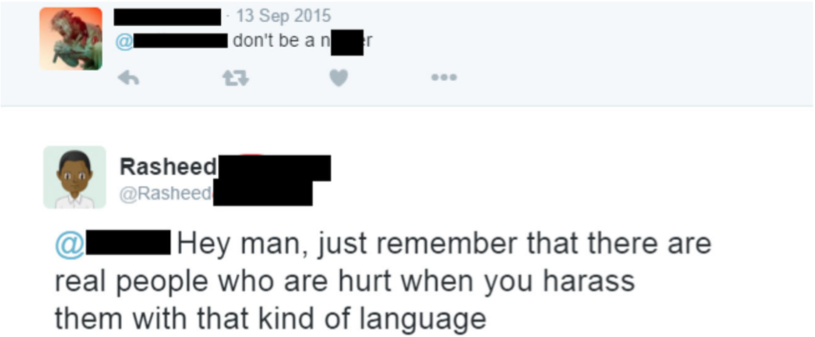
https://github.com/kmunger/Replication-Materials-for-Tweetment-Effects-on-the-Tweeted



[redacted] · 13 Sep 2015

@[redacted] don't be a n[redacted]r

See paper for more on the sampling procedure



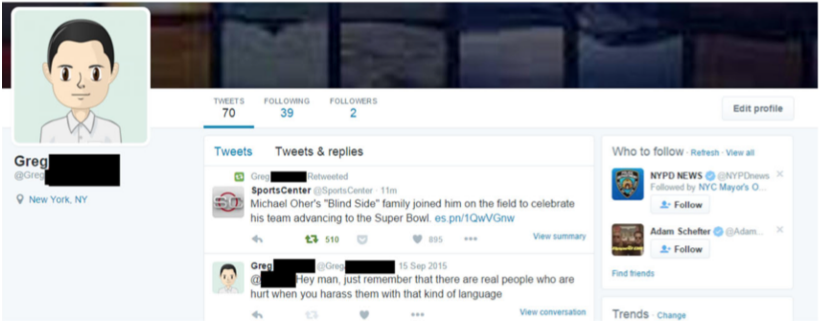
[redacted] · 13 Sep 2015

@[redacted] don't be a n[redacted]r

Rasheed[redacted]

@Rasheed

@[redacted] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language



Greg [redacted]
@Greg

New York, NY

TWEETS 70 FOLLOWING 39 FOLLOWERS 2

Greg [redacted] Retweeted

SportsCenter @SportsCenter · 11m
Michael Oher's "Blind Side" family joined him on the field to celebrate his team advancing to the Super Bowl. es.pn/1QwVGnw

510 895

Greg [redacted] @Greg [redacted] 15 Sep 2015
@[redacted] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language

NYPD NEWS @NYPDnews
Followed by NYC Mayor's O...

Adam Schefter @Adam...

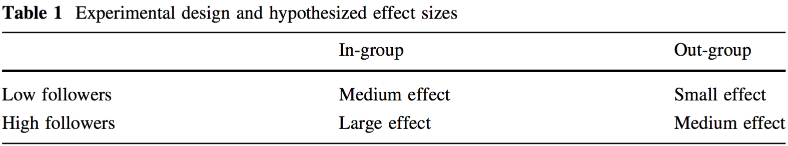


Table 1 Experimental design and hypothesized effect sizes

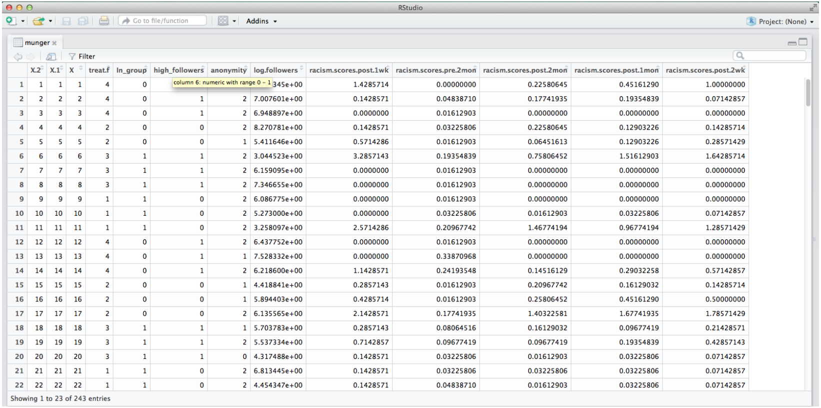
	In-group	Out-group
Low followers	Medium effect	Small effect
High followers	Large effect	Medium effect

2 x 2 design

Why do we need an experiment?

```r
munger <- read.csv("data/munger_tweetment_2017_data.csv")
summary(munger)
```

```
##       X.2             X.1              X            trea
##  Min.   :  1.0   Min.   :  1.0   Min.   :  1.0   Min.
##  1st Qu.: 61.5   1st Qu.: 61.5   1st Qu.: 61.5   1st Qu.
##  Median :122.0   Median :122.0   Median :122.0   Median
##  Mean   :122.0   Mean   :122.0   Mean   :122.1   Mean
##  3rd Qu.:182.5   3rd Qu.:182.5   3rd Qu.:182.5   3rd Qu.
##  Max.   :243.0   Max.   :243.0   Max.   :244.0   Max.
##
##     In_group      high_followers     anonymity     log.f
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000   1st Q
##  Median :0.0000   Median :0.0000   Median :2.000   Media
##  Mean   :0.4074   Mean   :0.4033   Mean   :1.547   Mean
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Q
##  Max.   :1.0000   Max.   :1.0000   Max.   :2.000   Max.
##
```



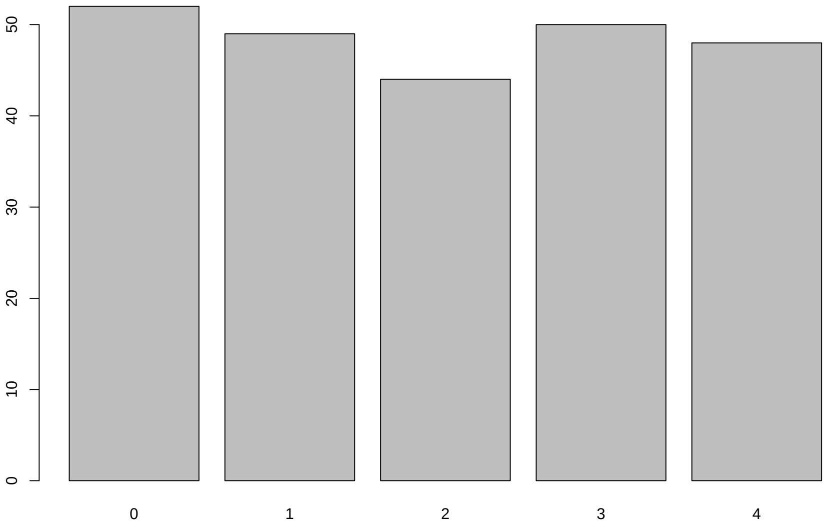
column 6: numeric with range 0 – 1

	X.2	X.1	X	treat.f	In_group	high_followers	anonymity	log.followers	racism.scores.post.1wk	racism.scores.pre.2mon	racism.scores.post.2mon	racism.scores.post.1mon	racism.scores.post.2wk
1	1	1	1	4	0			345e+00	1.4285714	0.00000000	0.22580645	0.45161290	1.00000000
2	2	2	2	4	0	1	2	7.007601e+00	0.1428571	0.04838710	0.17741935	0.19354839	0.07142857
3	3	3	3	4	0	1	2	6.948897e+00	0.0000000	0.01612903	0.00000000	0.00000000	0.00000000
4	4	4	4	2	0	0	2	8.270781e+00	0.1428571	0.03225806	0.22580645	0.12903226	0.14285714
5	5	5	5	2	0	0	1	5.411646e+00	0.5714286	0.01612903	0.06451613	0.12903226	0.28571429
6	6	6	6	3	1	1	2	3.044523e+00	3.2857143	0.19354839	0.75806452	1.51612903	1.64285714
7	7	7	7	3	1	1	2	6.159095e+00	0.0000000	0.01612903	0.00000000	0.00000000	0.00000000
8	8	8	8	3	1	1	2	7.346655e+00	0.0000000	0.01612903	0.00000000	0.00000000	0.00000000
9	9	9	9	1	1	0	2	6.086775e+00	0.0000000	0.01612903	0.00000000	0.00000000	0.00000000
10	10	10	10	1	1	0	2	5.273000e+00	0.0000000	0.03225806	0.01612903	0.03225806	0.07142857
11	11	11	11	1	1	0	2	3.258097e+00	2.5714286	0.20967742	1.46774194	0.96774194	1.28571429
12	12	12	12	4	0	1	2	6.437752e+00	0.0000000	0.01612903	0.00000000	0.00000000	0.00000000
13	13	13	13	4	0	1	1	7.528332e+00	0.0000000	0.33870968	0.00000000	0.00000000	0.00000000
14	14	14	14	4	0	1	2	6.218600e+00	1.1428571	0.24193548	0.14516129	0.29032258	0.57142857
15	15	15	15	2	0	0	1	4.418841e+00	0.2857143	0.01612903	0.20967742	0.16129032	0.14285714
16	16	16	16	2	0	0	1	5.894403e+00	0.4285714	0.01612903	0.25806452	0.45161290	0.50000000
17	17	17	17	2	0	0	2	6.135565e+00	2.1428571	0.17741935	1.40322581	1.67741935	1.78571429
18	18	18	18	3	1	1	1	5.703783e+00	0.2857143	0.08064516	0.16129032	0.09677419	0.21428571
19	19	19	19	3	1	1	2	5.537334e+00	0.7142857	0.09677419	0.09677419	0.19354839	0.42857143
20	20	20	20	3	1	1	0	4.317488e+00	0.1428571	0.03225806	0.01612903	0.03225806	0.07142857
21	21	21	21	1	1	0	2	6.813445e+00	0.1428571	0.03225806	0.03225806	0.03225806	0.07142857
22	22	22	22	1	1	0	2	4.454347e+00	0.1428571	0.04838710	0.01612903	0.03225806	0.07142857

Showing 1 to 23 of 243 entries

Data wrangling

```
barplot(table(munger$treat.f))
```

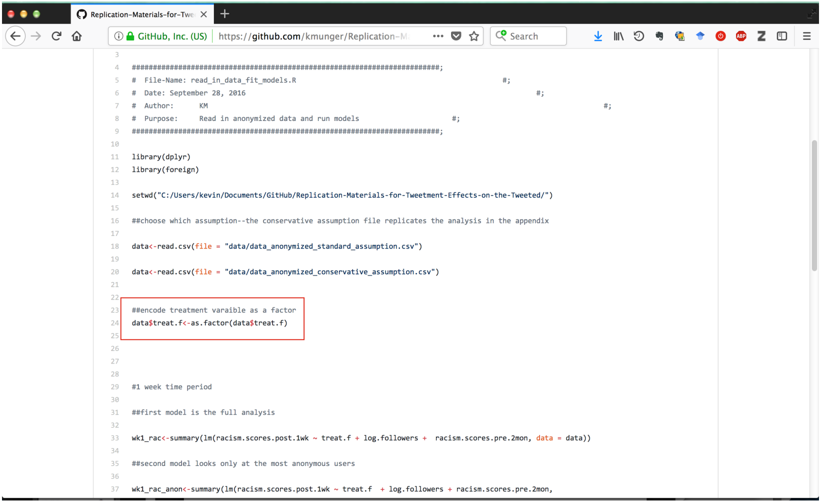


Data wrangling

```
str(munger$treat.f)
```

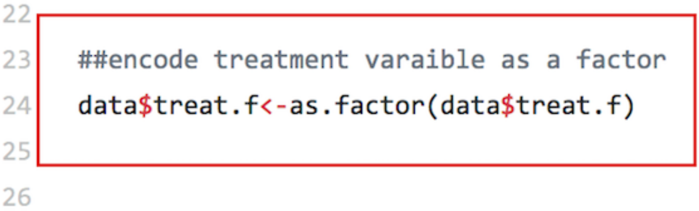
```
##  int [1:243] 4 4 4 2 2 3 3 3 1 1 ...
```

Data wrangling


Replication–Materials–for–Twe
GitHub, Inc. (US)
https://github.com/kmunger/Replication–M
Search

```r
#########################################################################;
#  File-Name: read_in_data_fit_models.R                                                  #;
#  Date: September 28, 2016                                                                      #;
#  Author:      KM                                                                                               #;
#  Purpose:     Read in anonymized data and run models                       #;
#########################################################################;

library(dplyr)
library(foreign)

setwd("C:/Users/kevin/Documents/GitHub/Replication-Materials-for-Tweetment-Effects-on-the-Tweeted/")

##choose which assumption--the conservative assumption file replicates the analysis in the appendix

data<-read.csv(file = "data/data_anonymized_standard_assumption.csv")

data<-read.csv(file = "data/data_anonymized_conservative_assumption.csv")


##encode treatment varaible as a factor
data$treat.f<-as.factor(data$treat.f)




#1 week time period

##first model is the full analysis

wk1_rac<-summary(lm(racism.scores.post.1wk ~ treat.f + log.followers +  racism.scores.pre.2mon, data = data))

##second model looks only at the most anonymous users

wk1_rac_anon<-summary(lm(racism.scores.post.1wk ~ treat.f  + log.followers + racism.scores.pre.2mon,
```

Data wrangling



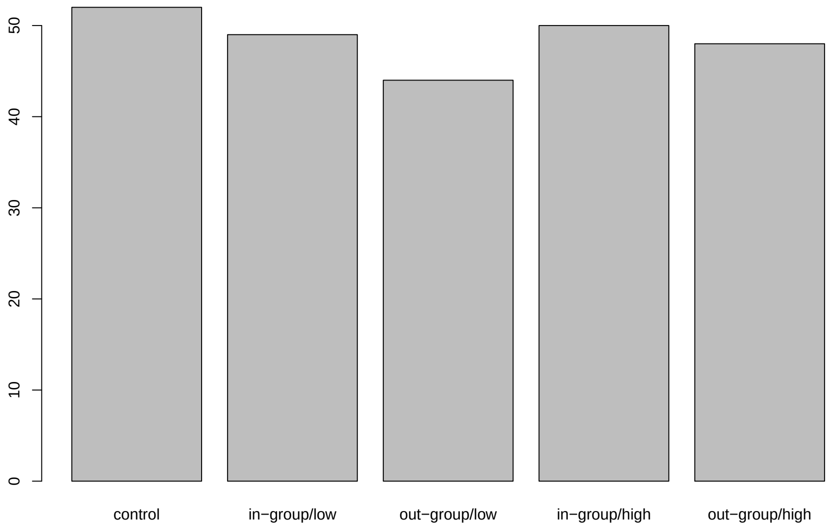
```
##encode treatment varaible as a factor
data$treat.f<-as.factor(data$treat.f)
```

Data wrangling

```r
munger$treat.f <- as.factor(munger$treat.f)
# 0 = control
# 1 = in-group, low followers
# 2 = out-group, low followers
# 3 = in-group, high followers
# 4 = out-group, high followers
levels(munger$treat.f) <- c("control",
                            "in-group/low",
                            "out-group/low",
                            "in-group/high",
                            "out-group/high")
```

Data wrangling

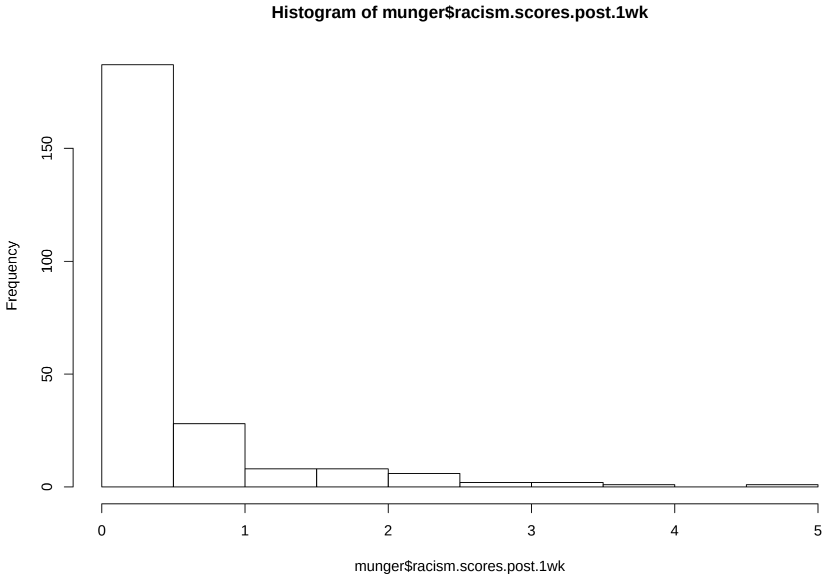
```r
barplot(table(munger$treat.f))
```



Data wrangling

Why am I not adding more informative labels?

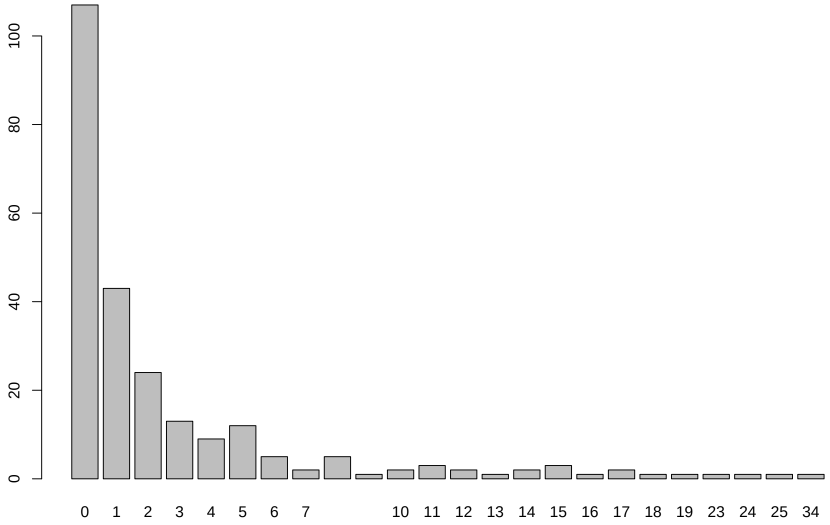
Data wrangling



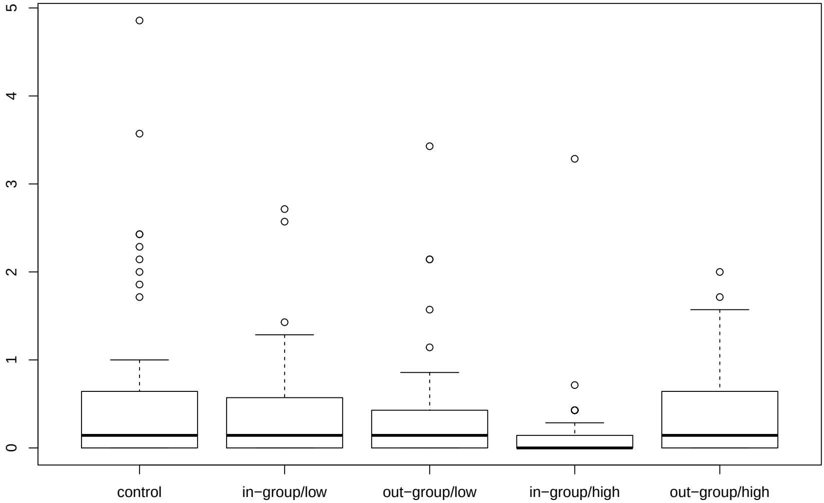
Data wrangling

> “Each panel shows the results of an OLS regression in which the dependent variable is the absolute number of instances of racists language during that period divided by the number of days in that time period.”

Data wrangling



Data wrangling

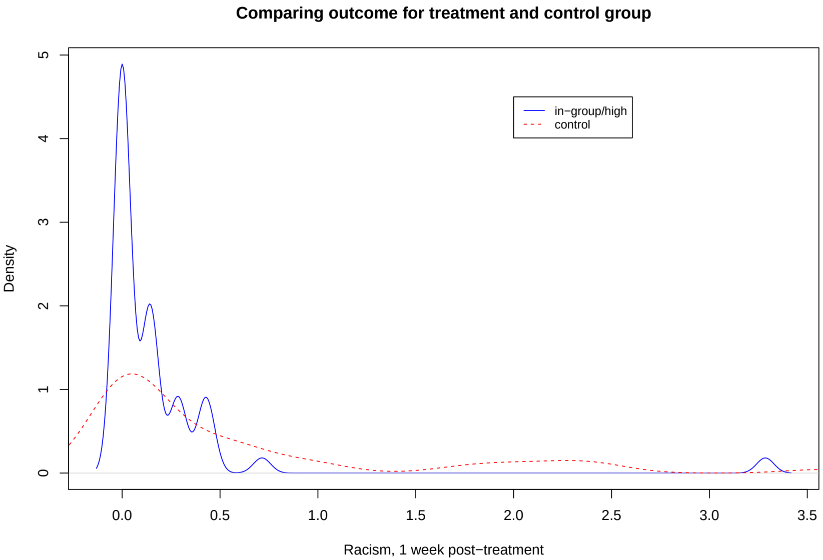


Data wrangling

```r
munger.twogroup <- subset(munger, subset = treat.f %in% c("
dim(munger.twogroup)
```

```
## [1] 102  13
```

Data wrangling



Intermission

What to make great plots without all the fiddling? Try ggplot2 at the COMPASS Workshop on December 7.

Intermission

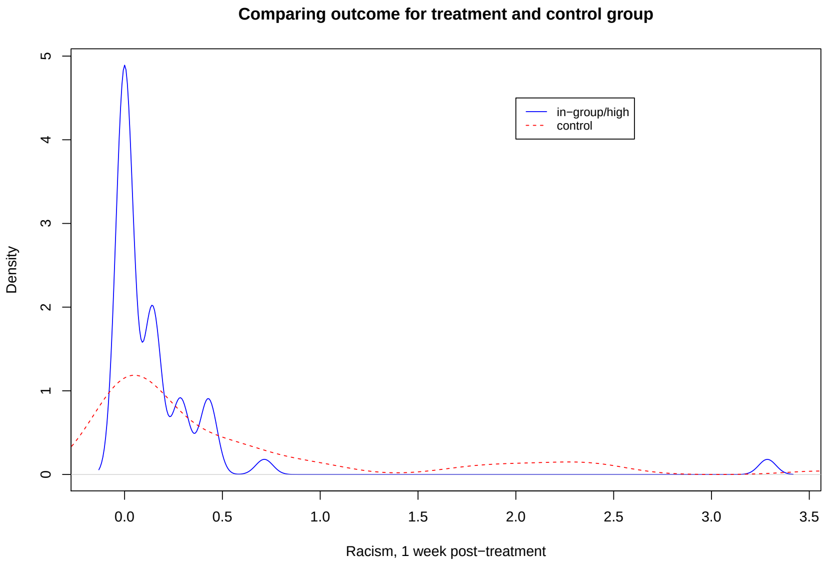
What to make great plots without all the fiddling? Try ggplot2 at the COMPASS Workshop on December 7. Can't wait that long?

Intermission

What to make great plots without all the fiddling? Try ggplot2 at the COMPASS Workshop on December 7. Can't wait that long?

PAID COURSE

Data Visualization with ggplot2 (Part 1)


Comparing outcome for treatment and control group
in−group/high
control
Density
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Racism, 1 week post−treatment

Difference-of-means approach

```r
y.treat <- mean(munger[munger$treat.f == "in-group/high", "
y.control <- mean(munger[munger$treat.f == "control", "raci
est.ate <- y.treat - y.control
print(paste("y.treat:", y.treat))
```

```
## [1] "y.treat: 0.182857142857143"
```

```r
print(paste("y.control:", y.control))
```

```
## [1] "y.control: 0.626373626373626"
```

```r
print(paste("est.ate:", est.ate))
```

```
## [1] "est.ate: -0.443516483516483"
```

The treated group created about 0.5 fewer racists post per day.

Difference-of-means approach

```r
n.treat <- sum(munger$treat.f == "in-group/high")
n.control <- sum(munger$treat.f == "control")
est.var.treat <- var(munger[munger$treat.f == "in-group/hig
est.var.control <- var(munger[munger$treat.f == "control",
est.se.ate <- sqrt(est.var.treat + est.var.control)

print(paste("est.se.ate:", est.se.ate))
```

```
## [1] "est.se.ate: 0.157452446428767"
```

Difference-of-means approach

```r
# 95% interval, rather than 1.96 you could use qnorm(0.975
lower.ci.95 <- est.ate - 1.96 * est.se.ate
upper.ci.95 <- est.ate + 1.96 * est.se.ate

print("Estimated 95 percent confidence interval:")
```

```
## [1] "Estimated 95 percent confidence interval:"
```

```r
print(c(lower.ci.95, upper.ci.95))
```

```
## [1] -0.7521233 -0.1349097
```

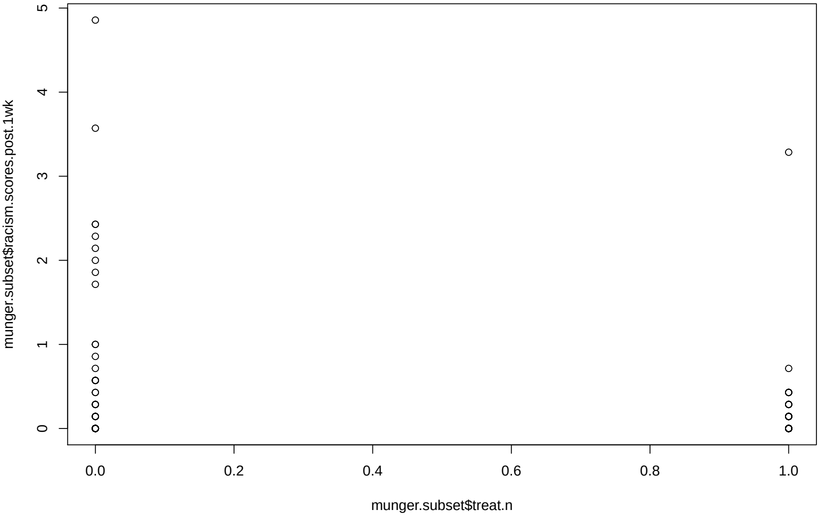
Regression approach

Let’s try that again with regression

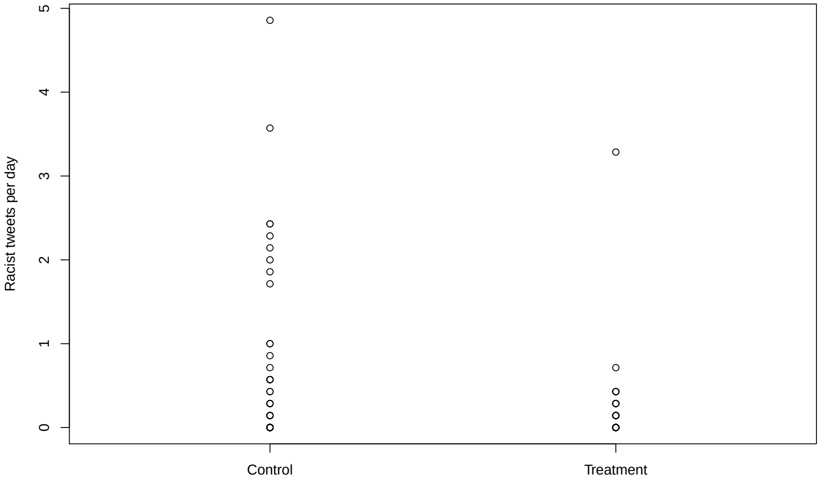
Regression approach, data wrangling

```
munger.subset <- subset(munger,
                        subset = treat.f %in%
                          c("control", "in-group/high"))
munger.subset$treat.n <- NA
cases.ih <- munger.subset$treat.f == "in-group/high"
munger.subset[cases.ih, "treat.n"] <- 1
cases.c <- munger.subset$treat.f == "control"
munger.subset[cases.c, "treat.n"] <- 0
```

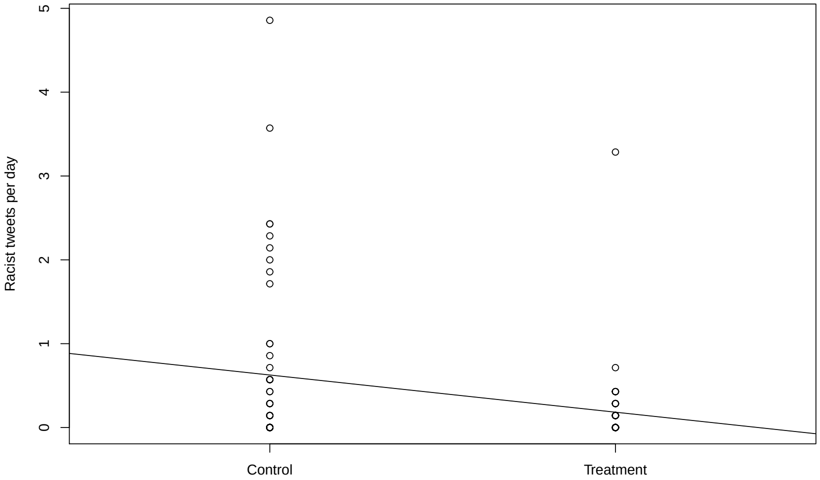
Regression approach



Regression approach



Regression approach



```r
fit <- lm(racism.scores.post.1wk ~ treat.n,
          data = munger.subset)
```

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ where

- $\hat{y}_i$ racist tweets per day

```r
fit <- lm(racism.scores.post.1wk ~ treat.n,
          data = munger.subset)
```

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ where

- $\hat{y}_i$ racist tweets per day
- x_i 1 if treatment, 0 if control

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ treat.n, data = mu
##
## Coefficients:
## (Intercept)      treat.n
##      0.6264      -0.4435
```

```
## [1] "y.treat: 0.182857142857143"
```

```
## [1] "y.control: 0.626373626373626"
```

The difference-of-means and the regression approach give us the same answer.[1] So why should we care about the regression approach?

[1] Technical note for interested folks: they can give slightly different estimated standard errors http://dx.doi.org/10.1016/j.spl.2011.10.024

The difference-of-means and the regression approach give us the same answer.[1] So why should we care about the regression approach?

It generalizes in interesting ways.

[1] Technical note for interested folks: they can give slightly different estimated standard errors http://dx.doi.org/10.1016/j.spl.2011.10.024

The difference-of-means and the regression approach give us the same answer.[1] So why should we care about the regression approach?

It generalizes in interesting ways.

- adjusting for pre-treament information

[1] Technical note for interested folks: they can give slightly different estimated standard errors http://dx.doi.org/10.1016/j.spl.2011.10.024

The difference-of-means and the regression approach give us the same answer.[1] So why should we care about the regression approach?

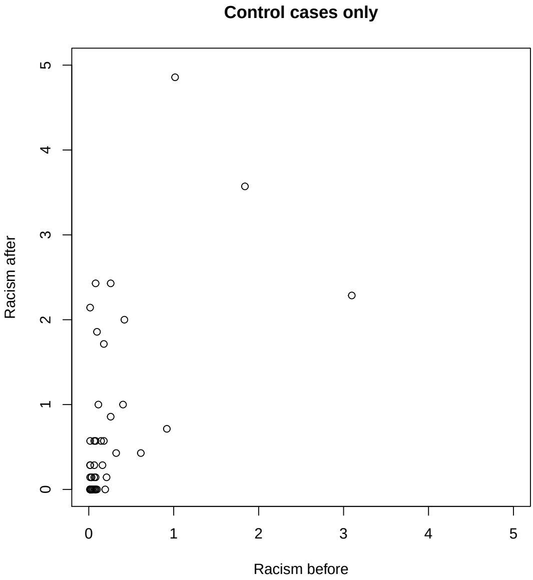
It generalizes in interesting ways.

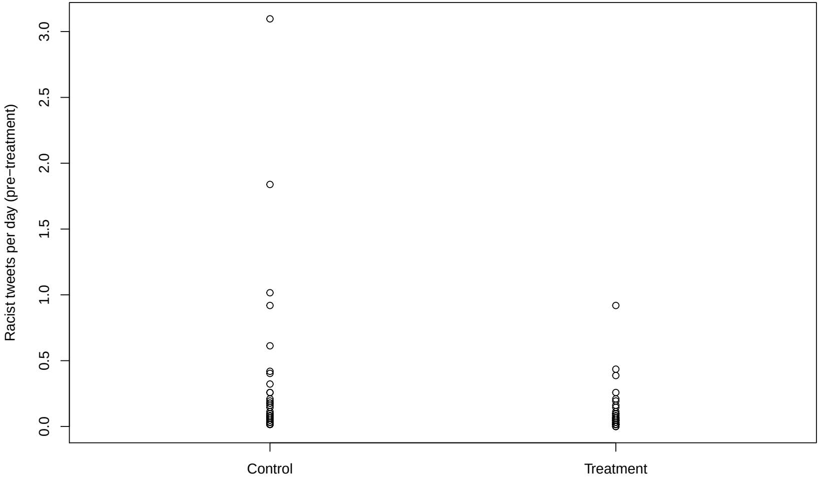
- adjusting for pre-treament information
- studying multiple treatments at the same time

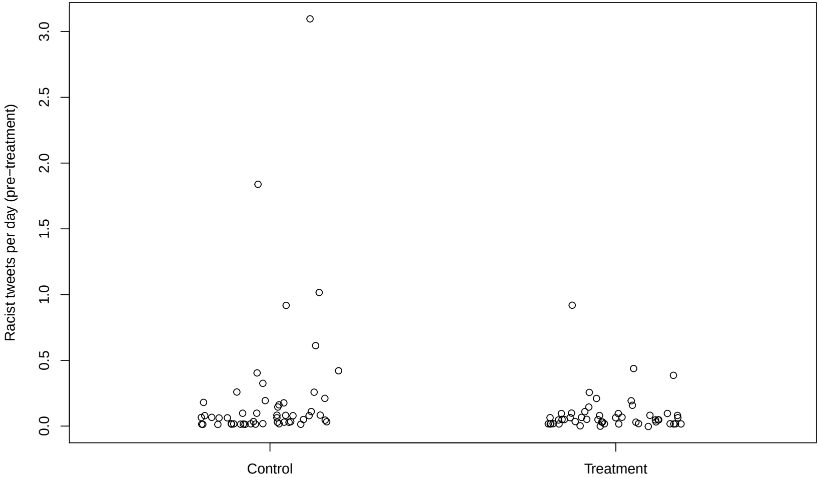
[1] Technical note for interested folks: they can give slightly different estimated standard errors http://dx.doi.org/10.1016/j.spl.2011.10.024

Adjusting for pre-treatment information

Being racist in the past predicts being racist in the future

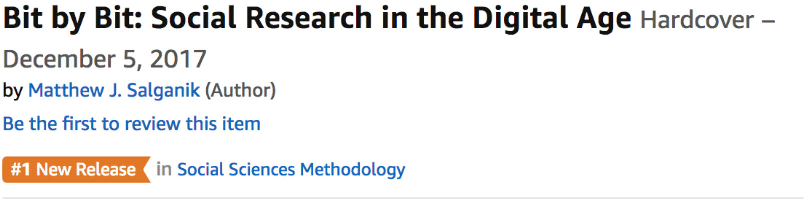



Racist tweets per day (pre−treatment)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Control
Treatment


Racist tweets per day (pre−treatment)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Control
Treatment

For more on including pre-treatment in the analysis of online field experiments:

- http://www.bitbybitbook.com/en/running-experiments/beyond-simple/



Bit by Bit: Social Research in the Digital Age Hardcover – December 5, 2017

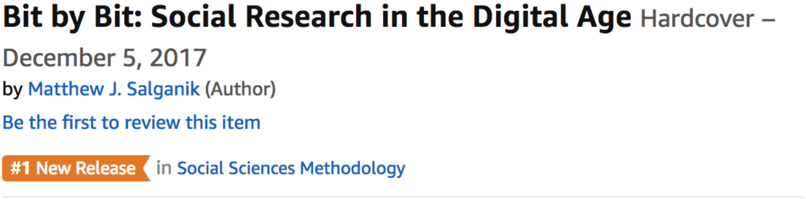
by Matthew J. Salganik (Author)

Be the first to review this item

#1 New Release in Social Sciences Methodology

For more on including pre-treatment in the analysis of online field experiments:

- http://www.bitbybitbook.com/en/running-experiments/beyond-simple/
- http://www.bitbybitbook.com/en/running-experiments/exp-advice/3rs/

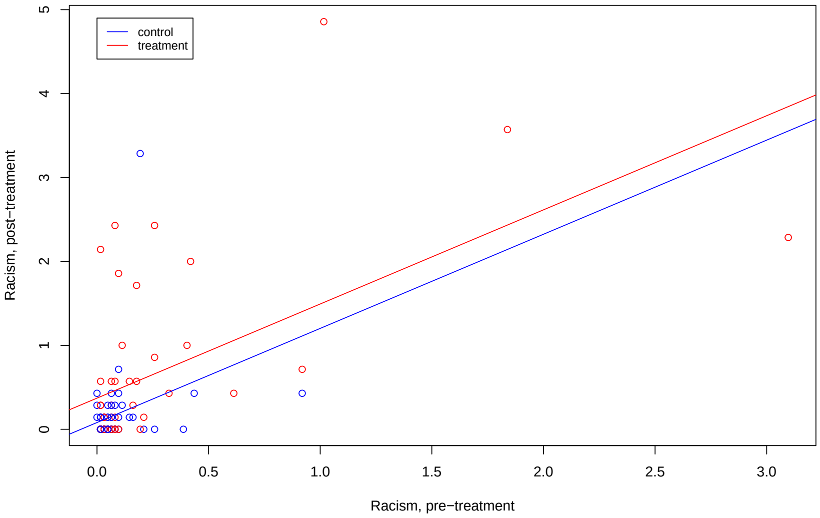


Bit by Bit: Social Research in the Digital Age Hardcover – December 5, 2017

by Matthew J. Salganik (Author)

Be the first to review this item

#1 New Release in Social Sciences Methodology


control
treatment
Racism, post−treatment
Racism, pre−treatment
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0

```
fit1 <- lm(racism.scores.post.1wk ~
             racism.scores.pre.2mon + treat.n,
           data = munger.subset)
```

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2}$ where

- $\hat{y}_i$ racist tweets per day, post-treatment

```
fit1 <- lm(racism.scores.post.1wk ~
             racism.scores.pre.2mon + treat.n,
           data = munger.subset)
```

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2}$ where

- $\hat{y}_i$ racist tweets per day, post-treatment
- $x_{i,1}$ racist tweets per day, pre-treatment

```r
fit1 <- lm(racism.scores.post.1wk ~
             racism.scores.pre.2mon + treat.n,
           data = munger.subset)
```

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2}$ where

- $\hat{y}_i$ racist tweets per day, post-treatment
- $x_{i,1}$ racist tweets per day, pre-treatment
- $x_{i,2}$ 1 if treatment, 0 if control

```r
lm(racism.scores.post.1wk ~ racism.scores.pre.2mon + treat.n,
           data = munger.subset)
```

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     treat.n, data = munger.subset)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                 treat.n
##                 0.3710                  1.1219                 -0.2909
```

```r
lm(racism.scores.post.1wk ~ treat.n,
          data = munger.subset)
```

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ treat.n, data = mu
##
## Coefficients:
## (Intercept)      treat.n
##      0.6264      -0.4435
```

- adjusting for pre-treament information

- adjusting for pre-treament information
- studying multiple treatments at the same time

Studying multiple treatments at the same time, data wrangling

Creating dummy variable

```
munger$control <- ifelse(munger$treat.f == "control",
                         1, 0)
munger$in.low <- ifelse(munger$treat.f == "in-group/low",
                         1, 0)
munger$out.low <- ifelse(munger$treat.f == "out-group/low",
                         1, 0)
munger$in.high <- ifelse(munger$treat.f == "in-group/high",
                         1, 0)
munger$out.high <- ifelse(munger$treat.f == "out-group/high
                         1, 0)
```

Studying multiple treatments at the same time, data wrangling

```
head(munger[,
            c("treat.f", "control", "in.low", "out.low", "in.high", "out.high")
     n = 10)
```

```
##            treat.f control in.low out.low in.high out.high
## 1  out-group/high       0      0       0       0        1
## 2  out-group/high       0      0       0       0        1
## 3  out-group/high       0      0       0       0        1
## 4   out-group/low       0      0       1       0        0
## 5   out-group/low       0      0       1       0        0
## 6   in-group/high       0      0       0       1        0
## 7   in-group/high       0      0       0       1        0
## 8   in-group/high       0      0       0       1        0
## 9    in-group/low       0      1       0       0        0
## 10   in-group/low       0      1       0       0        0
```

Studying multiple treatments at the same time

```r
lm(racism.scores.post.1wk ~ racism.scores.pre.2mon +
     in.low + out.low + in.high + out.high + control,
   data = munger)
```

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.
##     in.low + out.low + in.high + out.high + control, dat
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon
##                0.32525                 1.32264
##                out.low                 in.high
##               -0.01251                -0.26356
##                control
##                     NA
```

Why did this fail?

Broken model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 racist\_pre_i + \hat{\beta}_2 in\_low_i + \hat{\beta}_3 out\_low_i + \hat{\beta}_4 in\_high_i + \hat{\beta}_5 out\_high_i + \hat{\beta}_6 control_i$$

Why did this fail?

Broken model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 racist\_pre_i + \hat{\beta}_2 in\_low_i + \hat{\beta}_3 out\_low_i + \hat{\beta}_4 in\_high_i + \hat{\beta}_5 out\_high_i + \hat{\beta}_6 control_i$$

Better model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 racist\_pre_i + \hat{\beta}_2 in\_low_i + \hat{\beta}_3 out\_low_i + \hat{\beta}_4 in\_high_i + \hat{\beta}_5 out\_high_i$$

Why did this fail?

Broken model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 racist\_pre_i + \hat{\beta}_2 in\_low_i + \hat{\beta}_3 out\_low_i + \hat{\beta}_4 in\_high_i + \hat{\beta}_5 out\_high_i + \hat{\beta}_6 control_i$$

Better model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 racist\_pre_i + \hat{\beta}_2 in\_low_i + \hat{\beta}_3 out\_low_i + \hat{\beta}_4 in\_high_i + \hat{\beta}_5 out\_high_i$$

- Deeper explaination: Take Prof. Wasow's class

Why did this fail?

Broken model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 racist\_pre_i + \hat{\beta}_2 in\_low_i + \hat{\beta}_3 out\_low_i + \hat{\beta}_4 in\_high_i + \hat{\beta}_5 out\_high_i + \hat{\beta}_6 control_i$$

Better model:

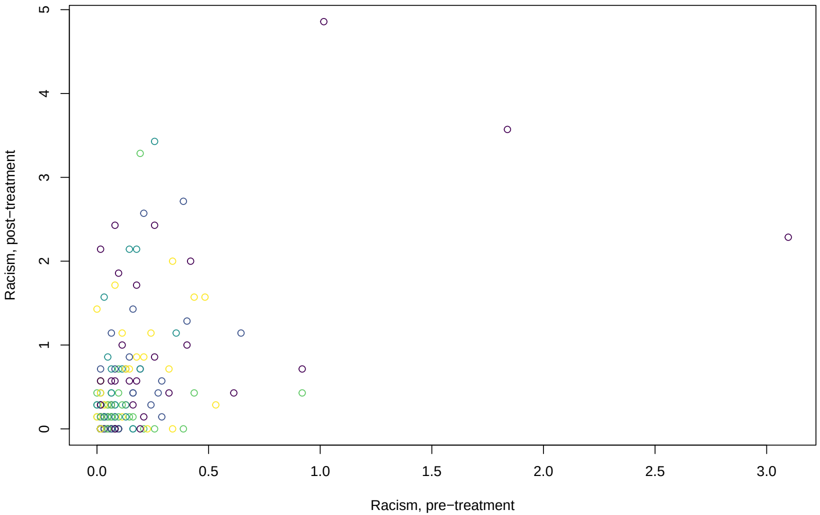
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 racist\_pre_i + \hat{\beta}_2 in\_low_i + \hat{\beta}_3 out\_low_i + \hat{\beta}_4 in\_high_i + \hat{\beta}_5 out\_high_i$$

- Deeper explaination: Take Prof. Wasow's class
- Can't wait: http://www.algosome.com/articles/dummy-variable-trap-regression.html

```
lm(racism.scores.post.1wk ~ racism.scores.pre.2mon +
    in.low + out.low + in.high + out.high,
   data = munger)
```

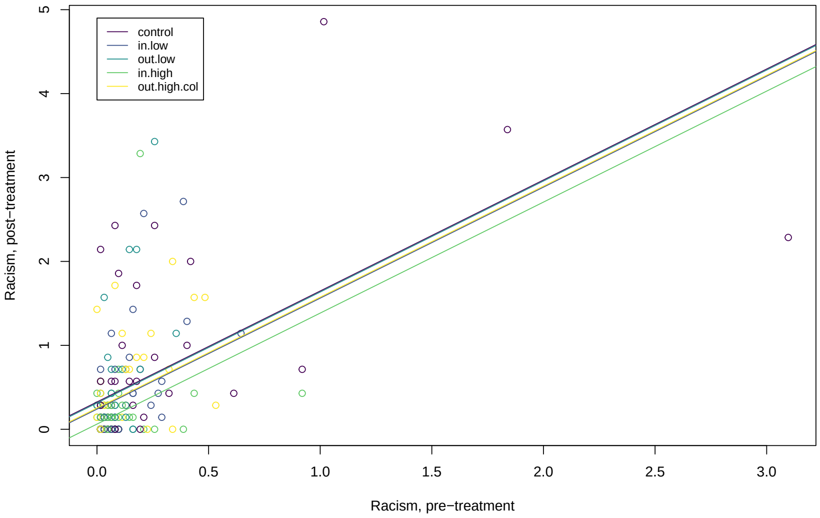
```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.
##     in.low + out.low + in.high + out.high, data = munger
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon
##                0.32525                 1.32264
##                out.low                 in.high
##               -0.01251                -0.26356
```

WHEN YOU HAVE DUMMY VARIABLES
DON'T FORGET AN
OMMITED CATEGORY


Racism, post−treatment
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Racism, pre−treatment

Better model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 racist\_pre_i + \hat{\beta}_2 in\_low_i + \hat{\beta}_3 out\_low_i + \hat{\beta}_4 in\_high_i + \hat{\beta}_5 out\_high_i$$


control
in.low
out.low
in.high
out.high.col
Racism, post−treatment
Racism, pre−treatment
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                  in.low
##                0.32525                 1.32264                -0.08529
##                out.low                 in.high                out.high
##               -0.01251                -0.26356                -0.07301
```

Which treatment is estimated to be the *most* effective?

1. in-group/low status

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##           (Intercept)  racism.scores.pre.2mon                  in.low
##               0.32525                 1.32264                -0.08529
##               out.low                 in.high                out.high
##              -0.01251                -0.26356                -0.07301
```

Which treatment is estimated to be the *most* effective?

1. in-group/low status
2. out-group/low status

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                  in.low
##                0.32525                 1.32264                -0.08529
##                out.low                 in.high                out.high
##               -0.01251                -0.26356                -0.07301
```

Which treatment is estimated to be the *most* effective?

1. in-group/low status
2. out-group/low status
3. in-group/high status

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                  in.low
##                0.32525                 1.32264                -0.08529
##                out.low                 in.high                out.high
##               -0.01251                -0.26356                -0.07301
```

Which treatment is estimated to be the *most* effective?

1. in-group/low status
2. out-group/low status
3. in-group/high status
4. out-group/high status

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                  in.low
##                0.32525                 1.32264                -0.08529
##                out.low                 in.high                out.high
##               -0.01251                -0.26356                -0.07301
```

Which treatment is estimated to be the *most* effective?

1. in-group/low status
2. out-group/low status
3. in-group/high status
4. out-group/high status

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                  in.low
##                0.32525                 1.32264                -0.08529
##                out.low                 in.high                out.high
##               -0.01251                -0.26356                -0.07301
```

Which treatment is estimated to be the *most* effective?

1. in-group/low status
2. out-group/low status
3. in-group/high status
4. out-group/high status

Answer: 3. in-group/high status

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                  in.low
##                0.32525                 1.32264                -0.08529
##                out.low                 in.high                out.high
##               -0.01251                -0.26356                -0.07301
```

Which treatment is estimated to be the least effective?

1. in-group/low status

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                   in.low
##                0.32525                 1.32264                 -0.08529
##                out.low                 in.high                 out.high
##               -0.01251                -0.26356                 -0.07301
```

Which treatment is estimated to be the least effective?

1. in-group/low status
2. out-group/low status

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                  in.low
##                0.32525                 1.32264                -0.08529
##                out.low                 in.high                out.high
##               -0.01251                -0.26356                -0.07301
```

Which treatment is estimated to be the least effective?

1. in-group/low status
2. out-group/low status
3. in-group/high status

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                  in.low
##                0.32525                 1.32264                -0.08529
##                out.low                 in.high                out.high
##               -0.01251                -0.26356                -0.07301
```

Which treatment is estimated to be the least effective?

1. in-group/low status
2. out-group/low status
3. in-group/high status
4. out-group/high status

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                  in.low
##                0.32525                 1.32264                -0.08529
##                out.low                 in.high                out.high
##               -0.01251                -0.26356                -0.07301
```

Which treatment is estimated to be the least effective?

1. in-group/low status
2. out-group/low status
3. in-group/high status
4. out-group/high status

Your turn

```
##
## Call:
## lm(formula = racism.scores.post.1wk ~ racism.scores.pre.2mon +
##     in.low + out.low + in.high + out.high, data = munger)
##
## Coefficients:
##            (Intercept)  racism.scores.pre.2mon                  in.low
##                0.32525                 1.32264                -0.08529
##                out.low                 in.high                out.high
##               -0.01251                -0.26356                -0.07301
```

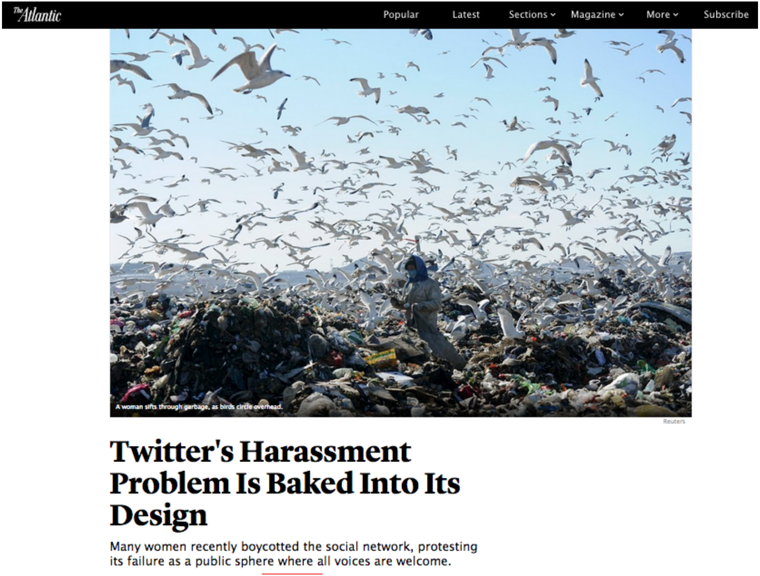
Which treatment is estimated to be the least effective?

1. in-group/low status
2. out-group/low status
3. in-group/high status
4. out-group/high status

Answer: 2. out-group/low status

Table 1 Experimental design and hypothesized effect sizes

	In-group	Out-group
Low followers	Medium effect	Small effect
High followers	Large effect	Medium effect



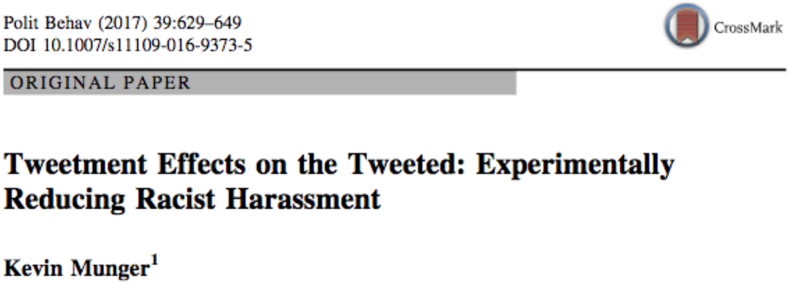
A woman sifts through garbage, as birds circle overhead.

Reuters

Twitter's Harassment Problem Is Baked Into Its Design

Many women recently boycotted the social network, protesting its failure as a public sphere where all voices are welcome.

https://www.theatlantic.com/technology/archive/2017/10/twitters-harassment-problem-is-baked-into-its-design/542952/



Polit Behav (2017) 39:629–649
DOI 10.1007/s11109-016-9373-5

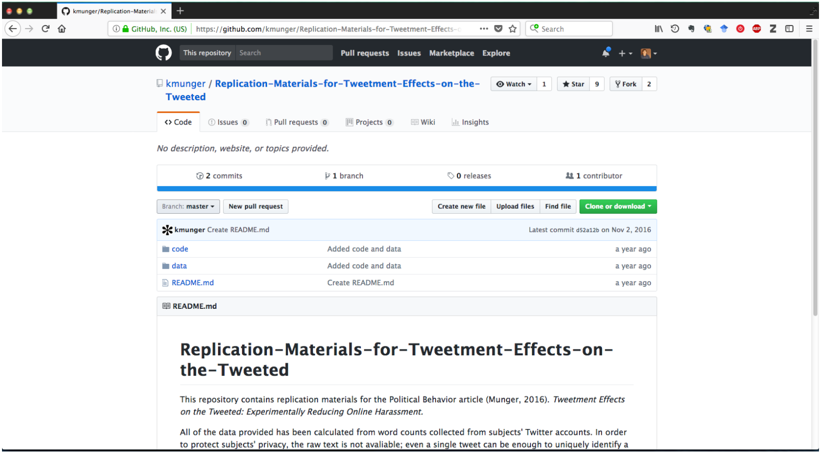
CrossMark

ORIGINAL PAPER

Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

Kevin Munger[1]

http://dx.doi.org/10.1007/s11109-016-9373-5



kmunger / **Replication-Materials-for-Tweetment-Effects-on-the-Tweeted**

No description, website, or topics provided.

2 commits | 1 branch | 0 releases | 1 contributor

kmunger Create README.md — Latest commit d52a12b on Nov 2, 2016

code	Added code and data	a year ago
data	Added code and data	a year ago
README.md	Create README.md	a year ago

README.md

Replication-Materials-for-Tweetment-Effects-on-the-Tweeted

This repository contains replication materials for the Political Behavior article (Munger, 2016). *Tweetment Effects on the Tweeted: Experimentally Reducing Online Harassment.*

All of the data provided has been calculated from word counts collected from subjects' Twitter accounts. In order to protect subjects' privacy, the raw text is not avaliable; even a single tweet can be enough to uniquely identify a

https://github.com/kmunger/
Replication-Materials-for-Tweetment-Effects-on-the-Tweeted}

Kevin Munger's next project: Experimentally Reducing Partisan Incivility on Twitter

- paper: http://kmunger.github.io/pdfs/jmp.pdf

Kevin Munger's next project: Experimentally Reducing Partisan Incivility on Twitter

- paper: http://kmunger.github.io/pdfs/jmp.pdf
- slides from talk at Twitter: http://kmunger.github.io/pdfs/munger_twitter_8_31.pdf

Goals for today

- See real data analyis workflow (with data wrangling)

Goals for today

- See real data analyis workflow (with data wrangling)
- Review difference-of-means

Goals for today

- See real data analyis workflow (with data wrangling)
- Review difference-of-means
- Show connection between difference-of-means and regression

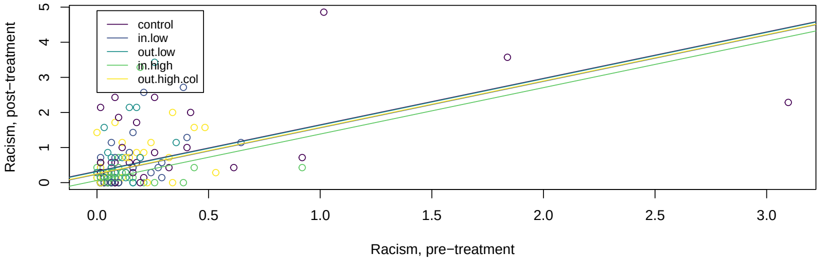
Goals for today

- See real data analyis workflow (with data wrangling)
- Review difference-of-means
- Show connection between difference-of-means and regression
- Explore multiple regression with continuous and dummy variables in equations, code, pictures, and words

Goals for today

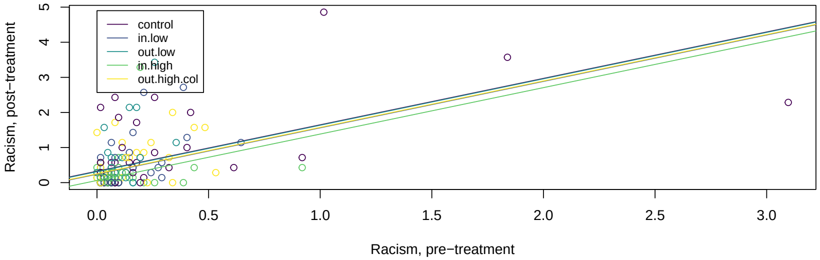
- See real data analyis workflow (with data wrangling)
- Review difference-of-means
- Show connection between difference-of-means and regression
- Explore multiple regression with continuous and dummy variables in equations, code, pictures, and words
- Learn something about Twitter

But there are open questions



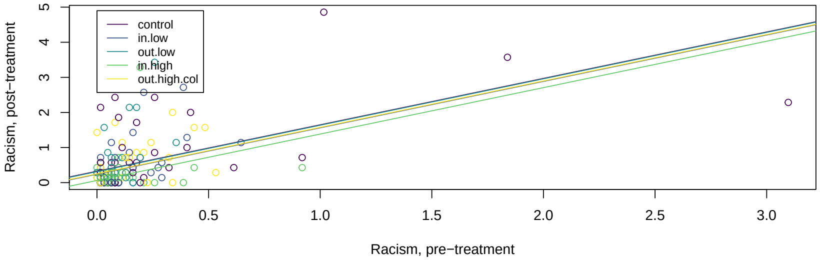
- What if the effect of the treatment varies based on the amount of racist speech pre-treatment?

But there are open questions



- What if the effect of the treatment varies based on the amount of racist speech pre-treatment?
- Are there more efficient ways to design an experiment like this?

But there are open questions



- What if the effect of the treatment varies based on the amount of racist speech pre-treatment?
- Are there more efficient ways to design an experiment like this?
- What about the ethics of all of this?

SOC 412: Designing Field Experiments at Scale

Online platforms, which monitor and intervene in the lives of billions of people, routinely host thousands of experiments to evaluate policies, test products, and contribute to theory in the social sciences. These experiments are also powerful tools to monitor injustice and govern human and algorithm behavior. How can we do field experiments at scale, reliably, and ethically?

SOC 412: Designing Field Experiments at Scale

By the end of the semester, you will be able to:

- Design, conduct, and interpret a novel online field experiment

Syllabus: http://natematias.com/courses/soc412/syllabus.html

SOC 412: Designing Field Experiments at Scale

By the end of the semester, you will be able to:

- Design, conduct, and interpret a novel online field experiment
- Write and critique a scholarly article reporting the results of the experiment

Syllabus: http://natematias.com/courses/soc412/syllabus.html

SOC 412: Designing Field Experiments at Scale

By the end of the semester, you will be able to:

- Design, conduct, and interpret a novel online field experiment
- Write and critique a scholarly article reporting the results of the experiment
- Design and analyze research from the perspective of rapid experimentation and reproduction in social science and industry

Syllabus: http://natematias.com/courses/soc412/syllabus.html

SOC 412: Designing Field Experiments at Scale

By the end of the semester, you will be able to:

- Design, conduct, and interpret a novel online field experiment
- Write and critique a scholarly article reporting the results of the experiment
- Design and analyze research from the perspective of rapid experimentation and reproduction in social science and industry
- Critically read, interpret, and imagine replications of the quantitative content of many field experiments in the social sciences

Syllabus: http://natematias.com/courses/soc412/syllabus.html

SOC 412: Designing Field Experiments at Scale

By the end of the semester, you will be able to:

- Design, conduct, and interpret a novel online field experiment
- Write and critique a scholarly article reporting the results of the experiment
- Design and analyze research from the perspective of rapid experimentation and reproduction in social science and industry
- Critically read, interpret, and imagine replications of the quantitative content of many field experiments in the social sciences
- Understand the kinds of knowledge that experiments bring to policy, product design, and theories in the social sciences, as well as their limitations

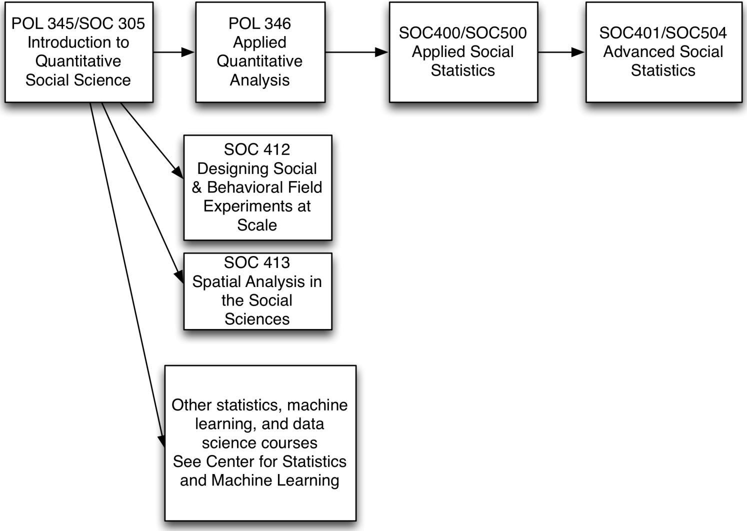
Syllabus: http://natematias.com/courses/soc412/syllabus.html

SOC 412: Designing Field Experiments at Scale

By the end of the semester, you will be able to:

- Design, conduct, and interpret a novel online field experiment
- Write and critique a scholarly article reporting the results of the experiment
- Design and analyze research from the perspective of rapid experimentation and reproduction in social science and industry
- Critically read, interpret, and imagine replications of the quantitative content of many field experiments in the social sciences
- Understand the kinds of knowledge that experiments bring to policy, product design, and theories in the social sciences, as well as their limitations
- Engage with debates on the ethics and politics of experiments in your own work

Syllabus: http://natematias.com/courses/soc412/syllabus.html



POL 345/SOC 305
Introduction to
Quantitative
Social Science

POL 346
Applied
Quantitative
Analysis

SOC400/SOC500
Applied Social
Statistics

SOC401/SOC504
Advanced Social
Statistics

SOC 412
Designing Social
& Behavioral Field
Experiments at
Scale

SOC 413
Spatial Analysis in
the Social
Sciences

Other statistics, machine
learning, and data
science courses
See Center for Statistics
and Machine Learning

Logistics

- QSS assignments due 24 hours before precept

Logistics

- QSS assignments due 24 hours before precept
- Pset 3 will be posted W 12/6 and due W 12/13

Logistics

- QSS assignments due 24 hours before precept
- Pset 3 will be posted W 12/6 and due W 12/13
- COMPASS workshop: Thurs, 11/30 Text Mining in R (Ethan)