W241 Final Project Data Analysis

R Markdown

```
# Load packages
library(foreign)
library(data.table)
library(knitr)
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
library(stringr)
```

```
# set an UDF to calculate robust standard error

get_robust_se <- function(mod) {
  sqrt(diag(vcovHC(mod)))
}
```

We can only download data in aggregate form from Facebook. Here, we are going to load the .csv file in and genearte individual records similar to what we were asked to do in PS4 problem 3.

```
# load the raw aggregated data downloaded from Facebook with gender and age
d_raw_ag <- fread("data_w241_final_project_ag.csv")
d_raw_ag <- as.data.frame(d_raw_ag)
# head(d_raw_ag)

# Transform the data by creating records from the aggregated records

ROW_COUNT = nrow(d_raw_ag)
reivew_image = c()
postive_review = c()
ad = c()
ad_desc = c()
age_group = c()
gender = c()
click = c()
```

```
for (i in 1:ROW_COUNT) {
  ad_set <- d_raw_ag[i,"Ad Set Name"]
  age <- d_raw_ag[i, "Age"]
  gen <- d_raw_ag[i, "Gender"]
  reach <- d_raw_ag[i, "Reach"]
  unique_clicks <- d_raw_ag[i, "Unique Link Clicks"]

  if (ad_set == "Ad Set for Ad A") {
    reivew_image <- append(reivew_image, rep(1, reach))
    postive_review <- append(postive_review, rep(1, reach))
    ad <- append(ad, rep("A", reach))
    ad_desc <- append(ad_desc, rep("Positve Review with Image", reach))
  } else if (ad_set == "Ad Set for Ad B") {
    reivew_image <- append(reivew_image, rep(0, reach))
    postive_review <- append(postive_review, rep(1, reach))
    ad <- append(ad, rep("B", reach))
    ad_desc <- append(ad_desc, rep("Positve Review without Image", reach))
  } else if (ad_set == "Ad Set for Ad C") {
    reivew_image <- append(reivew_image, rep(1, reach))
    postive_review <- append(postive_review, rep(0, reach))
    ad <- append(ad, rep("C", reach))
    ad_desc <- append(ad_desc, rep("Negative Review with Image", reach))
  } else {
    reivew_image <- append(reivew_image, rep(0, reach))
    postive_review <- append(postive_review, rep(0, reach))
    ad <- append(ad, rep("D", reach))
    ad_desc <- append(ad_desc, rep("Negative Review without Image", reach))
  }

  age_group <- append(age_group, rep(age, reach))
  gender <- append(gender, rep(gen, reach))

  if (is.na(unique_clicks)) {
    click <- append(click, rep(0, reach))
  }
  else {
    click <- append(click, rep(1, unique_clicks))
    click <- append(click, rep(0, reach - unique_clicks))
  }
}

d_ag <- data.table(
                     ad = ad,
                     ad_desc = ad_desc,
                     review_image = reivew_image,
                     positive_review = postive_review,
                     age_group = age_group,
                     gender = gender,
                     click = click)
```

```
rows <- sample(nrow(d_ag))
d_ag <- d_ag[rows, ]
kable(head(d_ag), caption = "Sample Records after Data Manipulation")
```

Table 1: Sample Records after Data Manipulation

ad	ad_desc	review_image	positive_review	age_group	gender	click
A	Positve Review with Image	1	1	45-54	male	0
B	Positve Review without Image	0	1	55-64	male	0
C	Negative Review with Image	1	0	35-44	male	0
C	Negative Review with Image	1	0	65+	male	0
B	Positve Review without Image	0	1	65+	male	0
B	Positve Review without Image	0	1	35-44	male	0

First Glance of the Result

Before we start analyzing the data, let's take a look result by treatment group first.

```
kable(d_ag[, .("Click Count" = sum(click),
               "Participant Count" = .N,
               "Click Rate" = paste(as.character(round(mean(click)*100,2)),"%")),
     keyby = .(Ad = str_replace(ad_desc, "\n", " "))],
     caption = "Result Overview")
```

Table 2: Result Overview

Ad	Click Count	Participant Count	Click Rate
Negative Review with Image	534	27188	1.96 %
Negative Review without Image	456	28691	1.59 %
Positve Review with Image	600	23552	2.55 %
Positve Review without Image	333	23449	1.42 %

There are two observations we have from the figure above: 1. Contrary to our believe, the presence of image didn't drive down the click rate of ad with negative review. That is, our hypothesis of image might attract attention and make people pay attention to review might not be valid. 2. We can see that the two ads with negative reviews had higher click rates than the ad with positive review without image. This observation suggests that the positivity of the review might not have any effect on people's interest in the product.

However, since the data was collected from a randomized sample of participants, there is inherent statistical uncertainty. We will have to do further analysis to calculate standard error on the effect of review image and review positivity in order to know the significance in the above observations.

Covariate Balance Check

We didn't have access to define or control the randomization process for A/B testing on Facebook. Therefore, we definitely would like to check the covariates' balance in different treatment groups before running regression analysis.

Quick look at the covariate balance across different treatment groups by gender and age group:

Look at the percentage of gender within different ads:

```
kable(d_ag[, round(prop.table(table(str_replace(ad_desc, "\n", " "), gender),1)*100,1)],
      caption = "Gender Composition(%) of Treatment Groups")
```

Table 3: Gender Composition(%) of Treatment Groups

	female	male	unknown
Negative Review with Image	10.0	89.2	0.8
Negative Review without Image	12.5	86.5	1.0
Positve Review with Image	6.8	92.3	0.9
Positve Review without Image	17.3	82.3	0.4

```
kable(d_ag[, round(prop.table(table(str_replace(ad_desc, "\n", " "), age_group),1)*100,1)],
      caption = "Age Composition(%) of Treatment Groups")
```

Table 4: Age Composition(%) of Treatment Groups

	18-24	25-34	35-44	45-54	55-64	65+
Negative Review with Image	4.4	16.9	26.2	25.9	16.7	9.8
Negative Review without Image	3.6	15.7	19.5	22.3	20.3	18.7
Positve Review with Image	3.0	12.9	20.3	29.3	23.0	11.5
Positve Review without Image	4.3	7.3	15.0	22.2	23.4	27.7

Looking at the table 2 and table 3, we noticed that the gender and age compositions of different treatment groups are not very even. Given our sample size of 102880, we would expect the compositions of treatment groups by gender and age to be very comparable. Therefore, it's important to note that we have significantly more males in the "Positive Review with Image" group and significantly more females in "Positive Review without Image" group. As for the age compositions, we have significantly more participants of the age group 65+ in the "Positive Review without Image" group. This is concerning because it is a sign that the randomization for this experiment might have failed. The power of the experimental method comes from the fact that the participants in different treatmnet groups were statistically equivalent and when these equivalent groups were exposed to different treatments and subsequent difference in results was observed, we can conclude that it was the treatment variations that CAUSED the results to be different. However, if participants in one treatment group already differ from those in other groups, we can't logically attribute the difference in the results observed to the treatment anymore since the difference could very well come from the participants. In our case, if gender and age turned to be assoicated with people's tendency to click an ad, we don't know wheter the difference of click rates we see if table 1 is due to the treatments or simply the difference between the participants in each the treatment groups.

Since the statistical equivalence between different treatment groups was the key to the vailidity of our experiment, we will perform formal tests to see whether covariates like gender and age were independent of the treatment variables, "review_image" and "positive_review".

Perform the test:

1. Use an F-test to check the whethter "gender" and "age_group" jointly have ability to predict treatment status compared with the null model which only has the bias term.

```
null_mod <- d_ag[ , lm(review_image ~ 1)]
full_mod <- d_ag[ , lm(review_image ~ as.factor(gender) + as.factor(age_group))]
anova_mod <- anova(full_mod, null_mod, test = 'F')
anova_mod
```

```
## Analysis of Variance Table
##
## Model 1: review_image ~ as.factor(gender) + as.factor(age_group)
## Model 2: review_image ~ 1
```

```
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1 102872 24780
## 2 102879 25715 -7   -935.67 554.92 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
null_mod <- d_ag[ , lm(positive_review ~ 1)]
full_mod <- d_ag[ , lm(positive_review ~ as.factor(gender) + as.factor(age_group))]
anova_mod <- anova(full_mod, null_mod, test = 'F')
anova_mod
```

```
## Analysis of Variance Table
##
## Model 1: positive_review ~ as.factor(gender) + as.factor(age_group)
## Model 2: positive_review ~ 1
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1 102872 25084
## 2 102879 25528 -7   -443.95 260.09 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both tests are highly significant which means the covariates jointly has the ability to predict the treatment status. Unfortunately, this means the randomization test failed. This means the execution of the experiment was not clean enough to make causal claim about our causal effect of interest. Despite this fact, we will still run a regression analysis on the data because we could demonstrate the above conclusion with a different vehicle as a practice and additional proof.

Regression Analysis

Since we have a 2x2 design, we will build a fully saturated model with 4 terms including the bias term. We will also also build three models with the gender and age_group covariates included separately and together.

```
mod_1 <- d_ag[, lm(click ~ review_image
                           + positive_review
                           + review_image * positive_review)]
# summary(mod_1)
mod_2 <- d_ag[, lm(click ~ review_image
                           + positive_review
                           + review_image * positive_review
                           + as.factor(gender))]
# summary(mod_2)
mod_3 <- d_ag[, lm(click ~ review_image
                           + positive_review
                           + review_image * positive_review
                           + as.factor(age_group))]
# summary(mod_3)
mod_4 <- d_ag[, lm(click ~ review_image
                           + positive_review
                           + review_image * positive_review
                           + as.factor(gender)
                           + as.factor(age_group))]

stargazer(
      mod_1,
      mod_2,
```

```
      mod_3,
      mod_4,
      se = list(get_robust_se(mod_1),
                get_robust_se(mod_2),
                get_robust_se(mod_3),
                get_robust_se(mod_4)),
      # type = 'text',
      title = "Regression Models",
      digits = 5,
      keep.stat = c("n", "rsq", "adj.rsq"),
      order = c("^review_image$", "^positive_review$", "review_image:positive_review", "Constant"))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Aug 10, 2020 - 02:01:30 PM

Looking at the first column in table 4, the naive interpretation is that the presence of a review image is predicted to increase the click rate by 0.38%. The model in column 1 also predicts the positivity of the text review to have no effect on the click rate and that the review image has an additional effect of 0.75% increase in click rate when the review is positive. However, the flaw in the randomization manifests when covariates that were supposed to have no effect on the point estimates of the treatment coefficients were included in regression. Though column 2 didn't show significant shifts in estimates of treatment effects likely due to gender having no association with the outcome variable, in column 3, we can see the estimated coefficients shifted for both review_image and positive_review. We think it's likely caused by the age_group variable being highly predictive of the outcome variable (similar results were obervered in column 4 where both gender and age_group were included in regression).

The significant shifts in the estimates of the treatment effect when covariates are included invalidate the finding in the fully saturated model in column 1. Calculating how much the treatment variable coefficients shift bewteen column 1 and column 3, we see review_image goes from 0.00375 to 0.0432 (15.2% increase), and postive_review goes from -0.00169 to -0.00268 (58.6% decrease).

This finding tells us the inclusion of covariates significantly change our estimated causal effect which should not happen with a successful randomiation because covariates should have been idependent of treatment variations. When we see treatment effect estiamtes shifting by the inclusion of covariates like this, we lose confidence in the results. Since we don't know whether there are other observed covariates which could also shift estimated treatment effect like age_group.

Appendix

Print the sample records of the raw data:

```
sample_records <- d_raw_ag[d_raw_ag$"Ad Set Name" %in%
                             c("Ad Set for Ad A"),
                           c("Ad Set Name", "Age", "Gender",
                             "Reach", "Unique Link Clicks")]

sample_records <- sample_records[order(sample_records$"Ad Set Name",
                                       sample_records$"Age",
                                       sample_records$"Gender"), ]

kable(sample_records,
      row.names = FALSE,
      caption = "Sample Records of Raw Data")
```

Table 5: Regression Models

	Dependent variable:			
	click			
	(1)	(2)	(3)	(4)
review_image	0.00375***	0.00379***	0.00432***	0.00434***
	(0.00112)	(0.00112)	(0.00113)	(0.00113)
positive_review	−0.00169	−0.00172	−0.00268**	−0.00267**
	(0.00107)	(0.00107)	(0.00108)	(0.00108)
review_image:positive_review	0.00753***	0.00759***	0.00760***	0.00761***
	(0.00170)	(0.00171)	(0.00171)	(0.00171)
Constant	0.01589***	0.01695***	0.00857***	0.00913***
	(0.00074)	(0.00138)	(0.00176)	(0.00207)
as.factor(gender)male		−0.00127		−0.00071
		(0.00132)		(0.00134)
as.factor(gender)unknown		0.00498		0.00551
		(0.00574)		(0.00575)
as.factor(age_group)25-34			0.00057	0.00060
			(0.00190)	(0.00190)
as.factor(age_group)35-44			0.00429**	0.00431**
			(0.00186)	(0.00187)
as.factor(age_group)45-54			0.01083***	0.01085***
			(0.00189)	(0.00190)
as.factor(age_group)55-64			0.01036***	0.01039***
			(0.00192)	(0.00192)
as.factor(age_group)65+			0.01007***	0.01000***
			(0.00197)	(0.00197)
Observations	102,880	102,880	102,880	102,880
R^2	0.00096	0.00098	0.00183	0.00185
Adjusted R^2	0.00093	0.00093	0.00175	0.00175
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6: Sample Records of Raw Data

Ad Set Name	Age	Gender	Reach	Unique Link Clicks
Ad Set for Ad A	18-24	female	16	2
Ad Set for Ad A	18-24	male	696	8
Ad Set for Ad A	18-24	unknown	0	NA
Ad Set for Ad A	25-34	female	215	6
Ad Set for Ad A	25-34	male	2808	44
Ad Set for Ad A	25-34	unknown	8	NA
Ad Set for Ad A	35-44	female	296	13
Ad Set for Ad A	35-44	male	4432	76
Ad Set for Ad A	35-44	unknown	53	NA
Ad Set for Ad A	45-54	female	528	12
Ad Set for Ad A	45-54	male	6304	168
Ad Set for Ad A	45-54	unknown	72	1
Ad Set for Ad A	55-64	female	434	14
Ad Set for Ad A	55-64	male	4968	152
Ad Set for Ad A	55-64	unknown	24	NA
Ad Set for Ad A	65+	female	120	9
Ad Set for Ad A	65+	male	2528	93
Ad Set for Ad A	65+	unknown	50	2