Customer Analytics HW 4

Qiutong Shi

Fri 8:30 – 11:30

1. is the treatment associated with a typically larger or lower number of shares?

```
> #simple linear regression of shares and treatment
> slr = lm(shares~if_videos, data=ds)
> summary(slr)

Call:
lm(formula = shares ~ if_videos, data = ds)

Residuals:
   Min     1Q Median     3Q    Max 
 -4309  -2310  -1691   -491 838990 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2891.47      73.58   39.30   <2e-16 ***
if_videos    1418.71     123.76   11.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11640 on 38710 degrees of freedom
Multiple R-squared:  0.003383,	Adjusted R-squared:  0.003358 
F-statistic: 131.4 on 1 and 38710 DF,  p-value: < 2.2e-16
```

The treatment is associated with a larger number of shares. According to the output of the linear regression, when treatment is present, number of shares will increase by 1419.

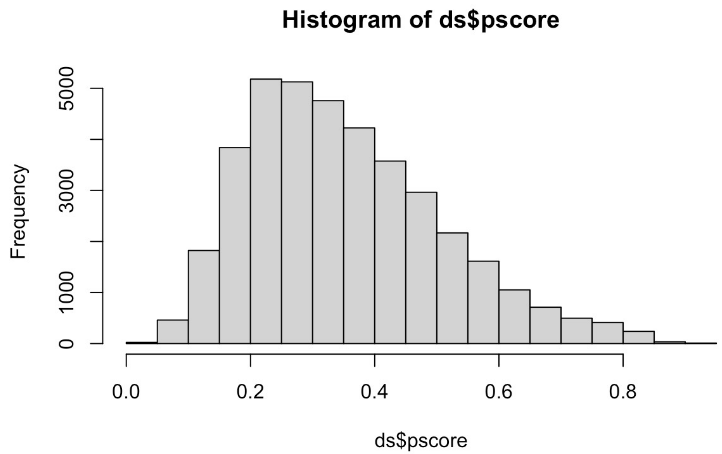
2.1 Evaluate the propensity score overlap between treated and non-treated subsamples.

From my knowledge, I think URL and timedelta have very little causational relationship with the outcome of number of shares. Therefore, it is not a confounding variable. Although I suspect that the number of videos can still affect the number of shares, I need to exclude it in my PS model to avoid it seperating the treatment indicator perfectly.

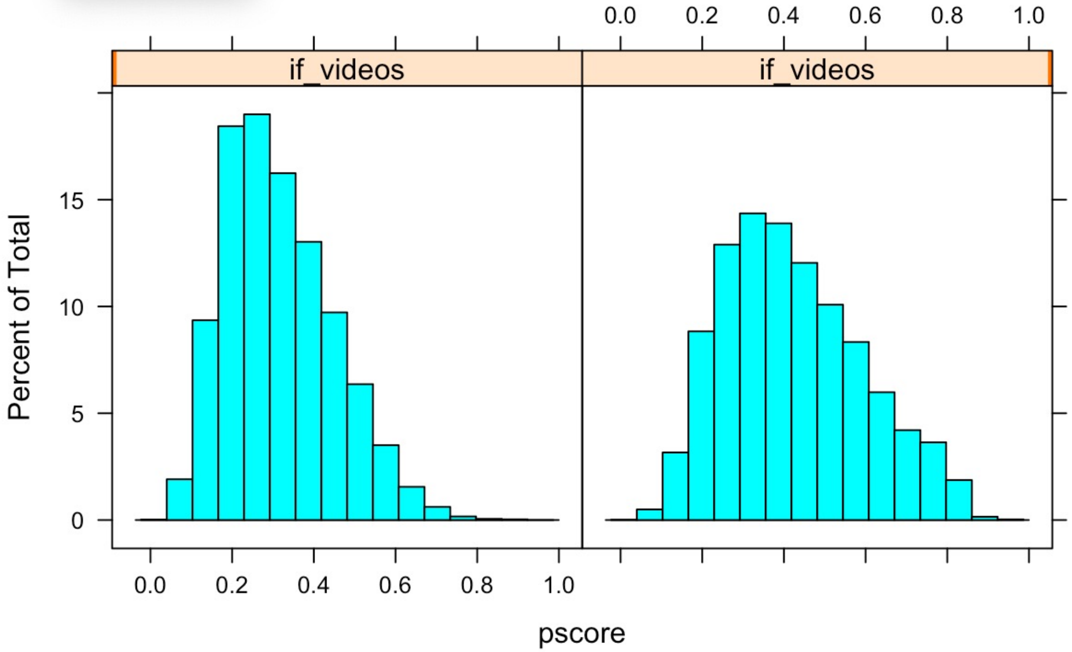
The link refers to my logic in PS model selecton:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1513192/

2.1 Evaluate the propensity score overlap between treated and non-treated subsamples.



We can observe that there is almos no observation beyond threshold = 0.8



2.2 Create a matched sample based on logistic propensity scores and in a way that accounts for overlap considerations

```r
# Perform matching with pre-focus on matching
matched <- matchit(if_videos~.-url-num_videos-timedelta-shares-pscore,method = "nearest",
                   data = ds[ds$pscore<=0.8,],distance='glm')
matched

# Create matched data set
ds_matched = match.data(matched)#convert matched object to macthed data set
dim(ds_matched)
table(ds_matched$if_videos)
```

29:46 (Top Level) ÷ R Script ÷

Console Terminal × Jobs ×

R 4.1.2 · ~/Desktop/luan/Customer Analytics/Week 5 matching analysis/HW4/

```
 - covariates: url, timedelta, n_tokens_title, n_tokens_content, n_unique_tokens, n_non_stop_unique_tokens, n
um_hrefs, num_self_hrefs, num_imgs, num_videos, average_token_length, num_keywords, kwshares_worst, kwshares_
best, kwshares_avg, self_reference_min_shares, self_reference_max_shares, self_reference_avg_sharess, global_
subjectivity, global_sentiment_polarity, global_rate_positive_words, global_rate_negative_words, rate_positiv
e_words, rate_negative_words, avg_positive_polarity, min_positive_polarity, max_positive_polarity, avg_negati
ve_polarity, min_negative_polarity, max_negative_polarity, title_subjectivity, title_sentiment_polarity, abs_
title_subjectivity, abs_title_sentiment_polarity, shares, category, weekday, pscore
> # Create matched data set
> ds_matched = match.data(matched)#convert matched object to macthed data set
> dim(ds_matched)
[1] 26850    42
> table(ds_matched$if_videos)

    0     1 
13425 13425 
>
```

I want to relate the treatment to the covariates on all variables except for url, num_videos, timedelta, shares, and pscores on ds, used in estimating the propensity score and for which balance is to be assessed. The method I choose is 1:1 near neighbor matching. I use glm in estimating the propensity score (basically asking matchit to replicate what I did for the previous question).

For an even number of control and test being matched, I have pre-filtered to focus on pscore<0.8 for enough masses.

2.3 has the matching procedure been successful?

The matching is successful. Before matching, the p-values of the means of the vairables of control and test group are small, indicating that they are statistically significant. This could damage the study of treatment.

After matching, most p-values of the means of the vairables of control and test group are large, indicating that they are not statistically significant.

The following slide would show the p-values side by side. Left-hand-side is before matching, right-hand-side is after matching.

2.3 has the matching procedure been successful?

```
> print(CreateTableOne(vars = xvars, data = ds, strata = "if_videos", smd = TRUE))
                                         Stratified by if_videos
                                          0                     1                     p      test
  n                                           25026                 13686
  n_tokens_title (mean (SD))                  10.21 (2.10)          10.69 (2.10)      <0.001
  n_tokens_content (mean (SD))               554.98 (449.36)       526.87 (501.64)    <0.001
  n_unique_tokens (mean (SD))                  0.56 (4.43)           0.53 (0.17)       0.429
  n_non_stop_unique_tokens (mean (SD))         0.70 (4.11)           0.67 (0.19)       0.286
  num_hrefs (mean (SD))                       11.04 (11.31)         10.61 (11.34)     <0.001
  num_self_hrefs (mean (SD))                   3.25 (4.06)           3.34 (3.52)       0.028
  num_imgs (mean (SD))                         4.68 (8.12)           4.20 (8.42)      <0.001
  average_token_length (mean (SD))             4.64 (0.63)           4.39 (1.11)      <0.001
  num_keywords (mean (SD))                     7.19 (1.94)           7.30 (1.84)      <0.001
  kwshares_worst (mean (SD))                 318.09 (510.46)       302.34 (768.52)     0.016
  kwshares_best (mean (SD))               236112.80 (121726.01) 271367.86 (133997.25) <0.001
  kwshares_avg (mean (SD))                  2969.13 (1164.23)     3259.51 (1453.11)   <0.001
  self_reference_min_shares (mean (SD))     4027.48 (21038.58)    4022.10 (17739.38)   0.980
  self_reference_max_shares (mean (SD))     8906.44 (35549.18)   12945.53 (49804.74)  <0.001
  self_reference_avg_sharess (mean (SD))    5873.04 (23925.79)    7433.21 (25319.44)  <0.001
  global_subjectivity (mean (SD))              0.44 (0.10)           0.44 (0.14)       0.018
  global_sentiment_polarity (mean (SD))        0.12 (0.09)           0.11 (0.10)      <0.001
  global_rate_positive_words (mean (SD))       0.04 (0.02)           0.04 (0.02)      <0.001
  global_rate_negative_words (mean (SD))       0.02 (0.01)           0.02 (0.01)      <0.001
  rate_positive_words (mean (SD))              0.70 (0.17)           0.65 (0.22)      <0.001
  rate_negative_words (mean (SD))              0.29 (0.15)           0.29 (0.17)       0.097
  avg_positive_polarity (mean (SD))            0.36 (0.09)           0.35 (0.12)      <0.001
  min_positive_polarity (mean (SD))            0.09 (0.07)           0.10 (0.08)      <0.001
  max_positive_polarity (mean (SD))            0.76 (0.23)           0.74 (0.28)      <0.001
  avg_negative_polarity (mean (SD))           -0.25 (0.12)          -0.27 (0.14)      <0.001
  min_negative_polarity (mean (SD))           -0.51 (0.28)          -0.53 (0.30)      <0.001
  max_negative_polarity (mean (SD))           -0.11 (0.09)          -0.11 (0.11)       0.003
  title_subjectivity (mean (SD))               0.27 (0.32)           0.31 (0.34)      <0.001
  title_sentiment_polarity (mean (SD))         0.07 (0.26)           0.07 (0.28)       0.760
  abs_title_subjectivity (mean (SD))           0.34 (0.19)           0.34 (0.19)       0.068
  abs_title_sentiment_polarity (mean (SD))     0.15 (0.22)           0.17 (0.24)      <0.001
```

```
> print(CreateTableOne(vars = xvars, data = ds_matched, strata = "if_videos", smd = TRUE))
                                         Stratified by if_videos
                                          0                     1                     p      test
  n                                           13425                 13425
  n_tokens_title (mean (SD))                  10.58 (2.12)          10.66 (2.10)       0.006
  n_tokens_content (mean (SD))               551.96 (456.51)       535.38 (501.68)     0.005
  n_unique_tokens (mean (SD))                  0.53 (0.13)           0.54 (0.16)      <0.001
  n_non_stop_unique_tokens (mean (SD))         0.67 (0.14)           0.68 (0.18)      <0.001
  num_hrefs (mean (SD))                       10.98 (11.40)         10.75 (11.21)      0.093
  num_self_hrefs (mean (SD))                   3.38 (4.30)           3.38 (3.42)       0.931
  num_imgs (mean (SD))                         4.59 (7.30)           4.28 (8.48)       0.001
  average_token_length (mean (SD))             4.55 (0.79)           4.45 (0.98)      <0.001
  num_keywords (mean (SD))                     7.28 (1.96)           7.30 (1.85)       0.400
  kwshares_worst (mean (SD))                 295.85 (473.33)       302.48 (760.17)     0.391
  kwshares_best (mean (SD))               254034.08 (118100.07) 266952.60 (130669.11) <0.001
  kwshares_avg (mean (SD))                  3101.51 (1225.33)     3215.61 (1352.63)   <0.001
  self_reference_min_shares (mean (SD))     3938.91 (17278.48)    4054.65 (17842.94)   0.589
  self_reference_max_shares (mean (SD))     9705.26 (37050.48)   11739.15 (39941.20)  <0.001
  self_reference_avg_sharess (mean (SD))    6097.96 (20873.86)    6961.62 (21740.94)   0.001
  global_subjectivity (mean (SD))              0.45 (0.11)           0.45 (0.13)       0.243
  global_sentiment_polarity (mean (SD))        0.12 (0.10)           0.11 (0.10)       0.082
  global_rate_positive_words (mean (SD))       0.04 (0.02)           0.04 (0.02)       0.563
  global_rate_negative_words (mean (SD))       0.02 (0.01)           0.02 (0.01)       0.082
  rate_positive_words (mean (SD))              0.68 (0.18)           0.66 (0.21)      <0.001
  rate_negative_words (mean (SD))              0.30 (0.15)           0.29 (0.16)       0.060
  avg_positive_polarity (mean (SD))            0.36 (0.10)           0.36 (0.11)       0.183
  min_positive_polarity (mean (SD))            0.10 (0.07)           0.10 (0.08)       0.429
  max_positive_polarity (mean (SD))            0.76 (0.24)           0.76 (0.26)       0.075
  avg_negative_polarity (mean (SD))           -0.27 (0.13)          -0.27 (0.14)       0.164
  min_negative_polarity (mean (SD))           -0.54 (0.29)          -0.54 (0.30)       0.849
  max_negative_polarity (mean (SD))           -0.11 (0.10)          -0.11 (0.10)       0.252
  title_subjectivity (mean (SD))               0.30 (0.33)           0.30 (0.34)       0.031
  title_sentiment_polarity (mean (SD))         0.07 (0.27)           0.07 (0.28)       0.331
  abs_title_subjectivity (mean (SD))           0.34 (0.19)           0.34 (0.19)       0.580
  abs_title_sentiment_polarity (mean (SD))     0.16 (0.23)           0.17 (0.23)       0.025
```

2.3 has the matching procedure been successful?

category (%)			<0.001
business	4893 (19.6)	1365 (10.0)	
entertainment	3152 (12.6)	2973 (21.7)	
lifestyle	1631 (6.5)	468 (3.4)	
socialmedia	1642 (6.6)	681 (5.0)	
tech	5275 (21.1)	2071 (15.1)	
world	8433 (33.7)	6128 (44.8)	
weekday (%)			<0.001
friday	3561 (14.2)	2008 (14.7)	
monday	4164 (16.6)	2313 (16.9)	
saturday	1677 (6.7)	724 (5.3)	
sunday	1791 (7.2)	853 (6.2)	
thursday	4619 (18.5)	2505 (18.3)	
tuesday	4541 (18.1)	2683 (19.6)	
wednesday	4673 (18.7)	2600 (19.0)	
if_videos (mean (SD))	0.00 (0.00)	1.00 (0.00)	<0.001

category (%)			0.001
business	1459 (10.9)	1357 (10.1)	
entertainment	2726 (20.3)	2915 (21.7)	
lifestyle	473 (3.5)	468 (3.5)	
socialmedia	727 (5.4)	679 (5.1)	
tech	2240 (16.7)	2068 (15.4)	
world	5800 (43.2)	5938 (44.2)	
weekday (%)			0.772
friday	1967 (14.7)	1964 (14.6)	
monday	2248 (16.7)	2272 (16.9)	
saturday	756 (5.6)	722 (5.4)	
sunday	907 (6.8)	850 (6.3)	
thursday	2454 (18.3)	2465 (18.4)	
tuesday	2555 (19.0)	2605 (19.4)	
wednesday	2538 (18.9)	2547 (19.0)	
if_videos (mean (SD))	0.00 (0.00)	1.00 (0.00)	<0.001

3.1 Based on your analysis above, provide a matching ATE estimate. Do videos increase the number of shares? By how much?

```
> # Estimate ATE
> summary(lm(shares ~ if_videos, data = ds_matched))

Call:
lm(formula = shares ~ if_videos, data = ds_matched)

Residuals:
   Min     1Q Median     3Q    Max
 -4272  -2773  -1973   -673 839027

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3034.2      115.2  26.330  < 2e-16 ***
if_videos     1238.8      163.0   7.601 3.03e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13350 on 26848 degrees of freedom
Multiple R-squared:  0.002148,	Adjusted R-squared:  0.00211
F-statistic: 57.78 on 1 and 26848 DF,  p-value: 3.027e-14
```

The average treatment effect is 1238.8. Videos increase the number of shares by 1239.

3.2 Provide a rationale that explains the disparity between the estimate of 3.a and 1.a

The estimate of 3a is smaller than 1a. The regression on a matched dataset shows that the impact of the treatment on number of shares might be smaller when other covariates are matched. Matching analysis helps alleviate potential model dependency by aligning the covariates.

3.3 what could then be the "fudge factor" (discussed in class) in this case?

A few fudge factors I could think of:

1. Time in a month the article is posted. It might be the case that people have no LTE or fast-speed data near the end of the month that affects whether they actually see an article with many videos and therefore share or not.
2. Pattern in sharing articles. People might feel uncomfortable sharing too many articles in a certain period of time to avoid flooding friends' inbox. For example, they could read 3 articles themselves and decide to share one.