Customer Analytics Assignment 4

Beiming Zhang

BU.450.760.K1.SP24

Light-colored text boxes contain code, and dark-colored text boxes contain output results.

Completely independent assignment.

Task 1 (a)

- Import and process data, converting some non-numeric variables into factors and constructing a treatment indicator(treat_ind).
- Build a linear regression model using the treatment indicator and shares.

```r
# Preparation of operating environment
library(tableone)
library(MatchIt)
library(lattice)
setwd(paste0(
    "/Users/velen/Documents/",
    "文稿-iCloud/Learning/JHU/Spring I/Customer Analytics/Class 5"
))
rm(list = ls())
ds <- read.csv("D5.2 Mashable.csv")

# Data cleaning
summary(ds)
table(ds$disclosed)
ds$category <- as.factor(ds$category)
ds$weekday <- as.factor(ds$weekday)

# Creat the treatment indicator
ds$treat_ind <- ifelse(ds$num_videos > 0, 1, 0)

# Task 1a
# Build the linear regression
model_t1 <- lm(shares ~ treat_ind, data = ds)
summary(model_t1)
```

Task 1 (a)

- In the linear regression results, the coefficient corresponding to the treatment indicator is significantly greater than 0, and the p-value is well below 0.01.
- Therefore, we can conclude that there is a positive correlation between the treatment indicator and shares, indicating that the treatment is associated with a typically larger number of shares.

```
> summary(model_t1)

Call:
lm(formula = shares ~ treat_ind, 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 ***
treat_ind    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
```

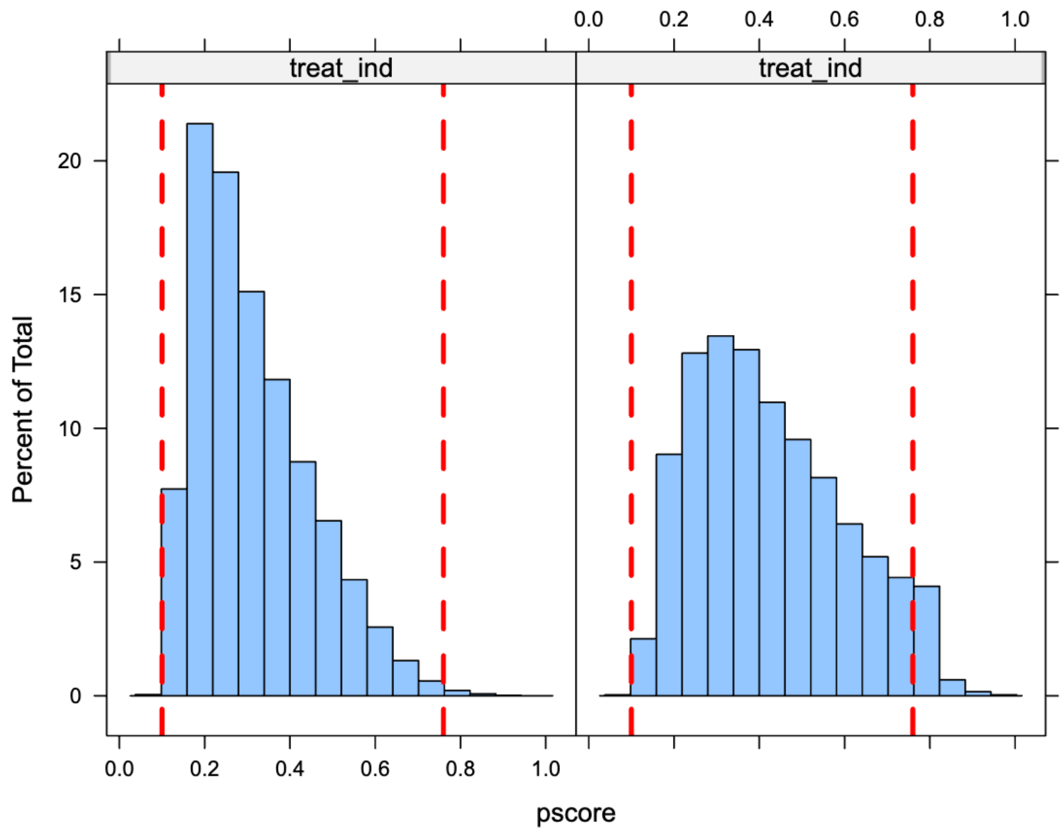
Task 2 (a)

- Remove irrelevant variables, retaining only appropriate covariates. Here, the “url”, "timedelta" and variables related to the independent and dependent variables were deleted.
- Generating a summary table that describes the baseline characteristics of the origin dataset and providing a statistical measure of how different the groups are from each other based on those variables.
- Calculate the propensity score using logistic regression and use a histogram to determine the overlap of the propensity score.

```r
# Task 2a
# Remove irrelevant covariates
drop_list <- c("url", "timedelta", "num_videos", "treat_ind")
xvars <- setdiff(names(ds), drop_list)

# Pre-match assessment of balance
table_unmatched <- CreateTableOne(
    vars = xvars, data = ds,
    strata = "treat_ind", smd = TRUE
)
print(table_unmatched)

# Calculate the pscore
y_x <- as.formula(paste("treat_ind ~", paste(xvars, collapse = " + ")))
model_t2 <- glm(y_x, family = binomial, data = ds)

# Assess pscore overlap
ds$pscore <- predict(model_t2, type = "response")
histogram(~ pscore | treat_ind, data = ds)

# Draw reference lines
histogram(~ pscore | treat_ind,
    data = ds,
    panel = function(x, ...) {
        panel.histogram(x, ...)
        panel.abline(v = c(0.1, 0.76), col = "red", lty = 2, lwd = 3)
    }
)
```

Task 2 (a)

- From the overlap diagram, the common overlap between the two is concentrated in the range of 0.1 <= n <= 0.76, indicating a relatively large overlap range.



Task 2 (b)

- Using the nearest neighbor matching method, perform the matching process based on the propensity score, and filter the matched dataset to ensure it only includes samples within the propensity score overlap region(0.1 <= n <= 0.76).
- In the result, there are 13462 values corresponding to 0 and 12518 values corresponding to 1, which is basically balanced.

```r
# Task 2b
# Perform matching
matched <- matchit(y_x, method = "nearest", data = ds)

# Filter based on the results of pscore overlap
ds_matched <- match.data(matched)
ds_matched <- ds_matched
    [ds_matched$pscore >= .1 & ds_matched$pscore <= .76, ]
table(ds_matched$treat_ind)
```

Task 2 (c)

- Construct a table to assess the sample in terms of covariate balance, comparing samples from two different datasets.
- After comparing the two (see the following 2 pages), I believe the matching procedure has been successful, as it effectively increased the differences in covariances corresponding to the samples. Although there are still some variables with small differences, this is due to the inherently small differences among the datasets themselves.

```r
# Task 2c
# Match assessment of balance
table_matched <- CreateTableOne(
    vars = xvars, data = ds_matched,
    strata = "treat_ind", smd = TRUE
)
print(table_unmatched)
print(table_matched)
```

Task 2 (c) (table_unmatched)

```
> print(table_unmatched)
                                         Stratified by treat_ind
                                          0
  n                                           25026
  n_tokens_title (mean (SD))                  10.21 (2.10)
  n_tokens_content (mean (SD))               554.98 (449.36)
  n_unique_tokens (mean (SD))                  0.56 (4.43)
  n_non_stop_unique_tokens (mean (SD))         0.70 (4.11)
  num_hrefs (mean (SD))                       11.04 (11.31)
  num_self_hrefs (mean (SD))                   3.25 (4.06)
  num_imgs (mean (SD))                         4.68 (8.12)
  average_token_length (mean (SD))             4.64 (0.63)
  num_keywords (mean (SD))                     7.19 (1.94)
  kwshares_worst (mean (SD))                 318.09 (510.46)
  kwshares_best (mean (SD))               236112.80 (121726.01)
  kwshares_avg (mean (SD))                  2969.13 (1164.23)
  self_reference_min_shares (mean (SD))     4027.48 (21038.58)
  self_reference_max_shares (mean (SD))     8906.44 (35549.18)
  self_reference_avg_sharess (mean (SD))    5873.04 (23925.79)
  global_subjectivity (mean (SD))              0.44 (0.10)
  global_sentiment_polarity (mean (SD))        0.12 (0.09)
  global_rate_positive_words (mean (SD))       0.04 (0.02)
  global_rate_negative_words (mean (SD))       0.02 (0.01)
  rate_positive_words (mean (SD))              0.70 (0.17)
  rate_negative_words (mean (SD))              0.29 (0.15)
  avg_positive_polarity (mean (SD))            0.36 (0.09)
  min_positive_polarity (mean (SD))            0.09 (0.07)
  max_positive_polarity (mean (SD))            0.76 (0.23)
  avg_negative_polarity (mean (SD))           -0.25 (0.12)
  min_negative_polarity (mean (SD))           -0.51 (0.28)
  max_negative_polarity (mean (SD))           -0.11 (0.09)
  title_subjectivity (mean (SD))               0.27 (0.32)
  title_sentiment_polarity (mean (SD))         0.07 (0.26)
  abs_title_subjectivity (mean (SD))           0.34 (0.19)
  abs_title_sentiment_polarity (mean (SD))     0.15 (0.22)
  shares (mean (SD))                        2891.47 (6524.88)
  category (%)
     business                                  4893 (19.6)
     entertainment                             3152 (12.6)
     lifestyle                                 1631 ( 6.5)
     socialmedia                               1642 ( 6.6)
     tech                                      5275 (21.1)
     world                                     8433 (33.7)
  weekday (%)
     friday                                    3561 (14.2)
     monday                                    4164 (16.6)
     saturday                                  1677 ( 6.7)
     sunday                                    1791 ( 7.2)
     thursday                                  4619 (18.5)
     tuesday                                   4541 (18.1)
     wednesday                                 4673 (18.7)
  pscore (mean (SD))                           0.31 (0.13)
```

```
                                          Stratified by treat_ind
                                           1                     p      test
  n                                            13686
  n_tokens_title (mean (SD))                   10.69 (2.10)      <0.001
  n_tokens_content (mean (SD))                526.87 (501.64)    <0.001
  n_unique_tokens (mean (SD))                   0.53 (0.17)       0.429
  n_non_stop_unique_tokens (mean (SD))          0.67 (0.19)       0.286
  num_hrefs (mean (SD))                        10.61 (11.34)     <0.001
  num_self_hrefs (mean (SD))                    3.34 (3.52)       0.028
  num_imgs (mean (SD))                          4.20 (8.42)      <0.001
  average_token_length (mean (SD))              4.39 (1.11)      <0.001
  num_keywords (mean (SD))                      7.30 (1.84)      <0.001
  kwshares_worst (mean (SD))                  302.34 (768.52)     0.016
  kwshares_best (mean (SD))                271367.86 (133997.25) <0.001
  kwshares_avg (mean (SD))                   3259.51 (1453.11)   <0.001
  self_reference_min_shares (mean (SD))      4022.10 (17739.38)   0.980
  self_reference_max_shares (mean (SD))     12945.53 (49804.74)  <0.001
  self_reference_avg_sharess (mean (SD))     7433.21 (25319.44)  <0.001
  global_subjectivity (mean (SD))               0.44 (0.14)       0.018
  global_sentiment_polarity (mean (SD))         0.11 (0.10)      <0.001
  global_rate_positive_words (mean (SD))        0.04 (0.02)      <0.001
  global_rate_negative_words (mean (SD))        0.02 (0.01)      <0.001
  rate_positive_words (mean (SD))               0.65 (0.22)      <0.001
  rate_negative_words (mean (SD))               0.29 (0.17)       0.097
  avg_positive_polarity (mean (SD))             0.35 (0.12)      <0.001
  min_positive_polarity (mean (SD))             0.10 (0.08)      <0.001
  max_positive_polarity (mean (SD))             0.74 (0.28)      <0.001
  avg_negative_polarity (mean (SD))            -0.27 (0.14)      <0.001
  min_negative_polarity (mean (SD))            -0.53 (0.30)      <0.001
  max_negative_polarity (mean (SD))            -0.11 (0.11)       0.003
  title_subjectivity (mean (SD))                0.31 (0.34)      <0.001
  title_sentiment_polarity (mean (SD))          0.07 (0.28)       0.760
  abs_title_subjectivity (mean (SD))            0.34 (0.19)       0.068
  abs_title_sentiment_polarity (mean (SD))      0.17 (0.24)      <0.001
  shares (mean (SD))                         4310.18 (17476.82)  <0.001
  category (%)                                                   <0.001
     business                                   1365 (10.0)
     entertainment                              2973 (21.7)
     lifestyle                                   468 ( 3.4)
     socialmedia                                 681 ( 5.0)
     tech                                       2071 (15.1)
     world                                      6128 (44.8)
  weekday (%)                                                    <0.001
     friday                                     2008 (14.7)
     monday                                     2313 (16.9)
     saturday                                    724 ( 5.3)
     sunday                                      853 ( 6.2)
     thursday                                   2505 (18.3)
     tuesday                                    2683 (19.6)
     wednesday                                  2600 (19.0)
  pscore (mean (SD))                            0.43 (0.18)      <0.001
```

Task 2 (c) (table_matched)

```
> print(table_matched)
                                          Stratified by treat_ind
                                          0
  n                                            13462
  n_tokens_title (mean (SD))                   10.56 (2.12)
  n_tokens_content (mean (SD))                554.22 (452.40)
  n_unique_tokens (mean (SD))                   0.53 (0.13)
  n_non_stop_unique_tokens (mean (SD))          0.67 (0.14)
  num_hrefs (mean (SD))                        10.99 (11.17)
  num_self_hrefs (mean (SD))                    3.37 (4.28)
  num_imgs (mean (SD))                          4.60 (7.08)
  average_token_length (mean (SD))              4.55 (0.78)
  num_keywords (mean (SD))                      7.29 (1.95)
  kwshares_worst (mean (SD))                  296.39 (551.56)
  kwshares_best (mean (SD))                254773.98 (114739.45)
  kwshares_avg (mean (SD))                   3104.75 (1220.46)
  self_reference_min_shares (mean (SD))      4098.66 (18995.66)
  self_reference_max_shares (mean (SD))      9217.70 (31513.51)
  self_reference_avg_sharess (mean (SD))     5996.02 (20855.77)
  global_subjectivity (mean (SD))               0.45 (0.11)
  global_sentiment_polarity (mean (SD))         0.12 (0.10)
  global_rate_positive_words (mean (SD))        0.04 (0.02)
  global_rate_negative_words (mean (SD))        0.02 (0.01)
  rate_positive_words (mean (SD))               0.68 (0.18)
  rate_negative_words (mean (SD))               0.30 (0.15)
  avg_positive_polarity (mean (SD))             0.36 (0.10)
  min_positive_polarity (mean (SD))             0.10 (0.07)
  max_positive_polarity (mean (SD))             0.76 (0.24)
  avg_negative_polarity (mean (SD))            -0.27 (0.13)
  min_negative_polarity (mean (SD))            -0.54 (0.29)
  max_negative_polarity (mean (SD))            -0.11 (0.10)
  title_subjectivity (mean (SD))                0.29 (0.33)
  title_sentiment_polarity (mean (SD))          0.07 (0.27)
  abs_title_subjectivity (mean (SD))            0.34 (0.19)
  abs_title_sentiment_polarity (mean (SD))      0.16 (0.23)
  shares (mean (SD))                         3039.33 (6158.58)
  category (%)
     business                                   1312 ( 9.7)
     entertainment                              2676 (19.9)
     lifestyle                                   484 ( 3.6)
     socialmedia                                 730 ( 5.4)
     tech                                       2275 (16.9)
     world                                      5985 (44.5)
  weekday (%)
     friday                                     1948 (14.5)
     monday                                     2257 (16.8)
     saturday                                    765 ( 5.7)
     sunday                                      880 ( 6.5)
     thursday                                   2480 (18.4)
     tuesday                                    2593 (19.3)
     wednesday                                  2539 (18.9)
  pscore (mean (SD))                            0.39 (0.12)
```

```
                                          Stratified by treat_ind
                                          1                     p      test
  n                                            12518
  n_tokens_title (mean (SD))                   10.62 (2.09)        0.011
  n_tokens_content (mean (SD))                554.50 (487.85)      0.962
  n_unique_tokens (mean (SD))                   0.55 (0.12)       <0.001
  n_non_stop_unique_tokens (mean (SD))          0.70 (0.13)       <0.001
  num_hrefs (mean (SD))                        11.14 (11.17)       0.271
  num_self_hrefs (mean (SD))                    3.48 (3.39)        0.022
  num_imgs (mean (SD))                          4.36 (8.05)        0.011
  average_token_length (mean (SD))              4.60 (0.56)       <0.001
  num_keywords (mean (SD))                      7.29 (1.87)        0.954
  kwshares_worst (mean (SD))                  301.08 (684.95)      0.543
  kwshares_best (mean (SD))                257748.39 (125972.64)   0.046
  kwshares_avg (mean (SD))                   3132.25 (1266.04)     0.075
  self_reference_min_shares (mean (SD))      4136.02 (18164.96)    0.871
  self_reference_max_shares (mean (SD))     10691.08 (31289.92)   <0.001
  self_reference_avg_sharess (mean (SD))     6657.88 (19952.44)    0.009
  global_subjectivity (mean (SD))               0.46 (0.10)       <0.001
  global_sentiment_polarity (mean (SD))         0.12 (0.10)        0.040
  global_rate_positive_words (mean (SD))        0.04 (0.02)       <0.001
  global_rate_negative_words (mean (SD))        0.02 (0.01)       <0.001
  rate_positive_words (mean (SD))               0.69 (0.17)       <0.001
  rate_negative_words (mean (SD))               0.30 (0.16)        0.016
  avg_positive_polarity (mean (SD))             0.37 (0.10)       <0.001
  min_positive_polarity (mean (SD))             0.10 (0.07)        0.001
  max_positive_polarity (mean (SD))             0.78 (0.23)       <0.001
  avg_negative_polarity (mean (SD))            -0.28 (0.13)       <0.001
  min_negative_polarity (mean (SD))            -0.56 (0.29)        0.001
  max_negative_polarity (mean (SD))            -0.11 (0.10)        0.005
  title_subjectivity (mean (SD))                0.30 (0.33)        0.076
  title_sentiment_polarity (mean (SD))          0.07 (0.28)        0.334
  abs_title_subjectivity (mean (SD))            0.34 (0.19)        0.925
  abs_title_sentiment_polarity (mean (SD))      0.17 (0.23)        0.039
  shares (mean (SD))                         3353.21 (6213.86)    <0.001
  category (%)                                                     0.002
     business                                   1305 (10.4)
     entertainment                              2704 (21.6)
     lifestyle                                   462 ( 3.7)
     socialmedia                                 674 ( 5.4)
     tech                                       2056 (16.4)
     world                                      5317 (42.5)
  weekday (%)                                                      0.995
     friday                                     1831 (14.6)
     monday                                     2118 (16.9)
     saturday                                    697 ( 5.6)
     sunday                                      799 ( 6.4)
     thursday                                   2296 (18.3)
     tuesday                                    2424 (19.4)
     wednesday                                  2353 (18.8)
  pscore (mean (SD))                            0.40 (0.14)       <0.001
```

Task 3 (a)

- Reconstruct the linear regression model using the matched dataset, where the coefficient of treat_ind represents the estimated ATE of videos on the number of shares.

```
# Task 3a
# Build the linear regression based on ds_matched
model_t3 <- lm(shares ~ treat_ind, data = ds_matched)
summary(model_t3)
```

Task 3 (a)

- Based on the regression results, with the coefficient of treat_ind being positive, we can conclude that videos indeed increase the number of shares, with an estimated increase of about 313.84 shares.

```
> summary(model_t3)

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

Residuals:
   Min     1Q Median     3Q    Max
 -3347  -2248  -1734   -348 141366

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3034.49      53.09  57.154  < 2e-16 ***
treat_ind     313.84      76.48   4.104 4.08e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6174 on 26096 degrees of freedom
Multiple R-squared:  0.0006449, Adjusted R-squared:  0.0006066
F-statistic: 16.84 on 1 and 26096 DF,  p-value: 4.081e-05
```

Task 3 (b)

- The data in 1.a was unprocessed, and after reprocessing with the propensity score, the samples associated with treat_ind became more balanced, reducing selection bias. This leads to a smoother and more objective overall sample distribution.
- Matching makes the groups more similar in terms of important covariates, thereby reducing the bias these covariates might introduce.
- At the same time, we removed some outliers, which will not be included in the model in 3.a. This also leads to a smaller coefficient in 3.a compared to 1.a.

Task 3 (c)

- The "fudge factor" in this case may include the following:
    1. The relevance between video content and article content. If video content is highly relevant to the article and complements it, then the likelihood of shares may increase, as it reflects the overall quality of the article.
    2. The total length of the video. The length of a video generally represents the amount of information in the article; if a video is longer, it typically contains more information and is more likely to be considered "valuable" and shared.
    3. The platform where the video is located. Videos on X are usually shorter and often hastily shot, with lower video quality; whereas videos on YouTube might have higher production quality and therefore more opportunities for shares.

Thank you

Beiming Zhang

BU.450.760.K1.SP24