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)
options(width = 120)
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
```

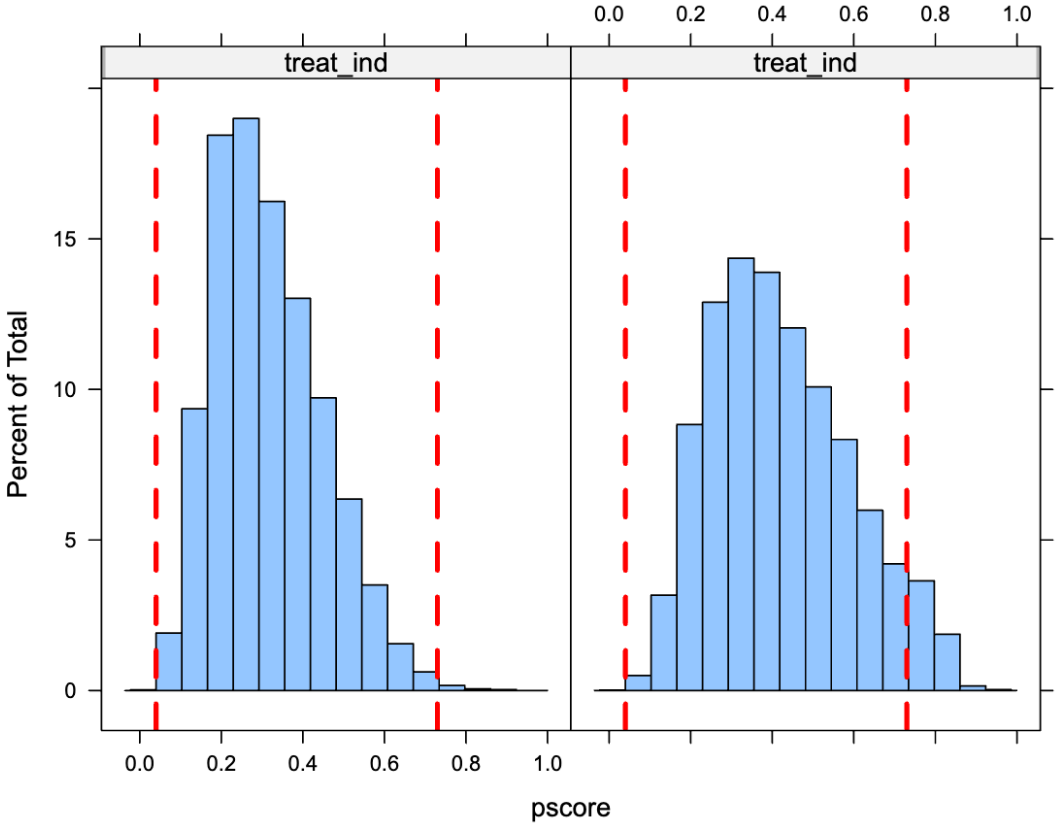
Task 2 (a)

- 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", "shares", "treat_ind", "num_videos")
xvars <- setdiff(names(ds), drop_list)

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

# 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(.04, .73), col = "red", lty = 2, lwd = 3)
    }
)
```

Task 2 (a)

```
> print(table_unmatched, smd = TRUE)
                                        Stratified by treat_ind
                                         0                    1                    p      test SMD
  n                                          25026                13686
  n_tokens_title (mean (SD))                 10.21 (2.10)         10.69 (2.10)     <0.001       0.226
  n_tokens_content (mean (SD))              554.98 (449.36)      526.87 (501.64)   <0.001       0.059
  n_unique_tokens (mean (SD))                 0.56 (4.43)          0.53 (0.17)      0.429       0.010
  n_non_stop_unique_tokens (mean (SD))        0.70 (4.11)          0.67 (0.19)      0.286       0.013
  num_hrefs (mean (SD))                      11.04 (11.31)        10.61 (11.34)    <0.001       0.038
  num_self_hrefs (mean (SD))                  3.25 (4.06)          3.34 (3.52)      0.028       0.024
  num_imgs (mean (SD))                        4.68 (8.12)          4.20 (8.42)     <0.001       0.059
  average_token_length (mean (SD))            4.64 (0.63)          4.39 (1.11)     <0.001       0.283
  num_keywords (mean (SD))                    7.19 (1.94)          7.30 (1.84)     <0.001       0.057
  kwshares_worst (mean (SD))                318.09 (510.46)      302.34 (768.52)    0.016       0.024
  kwshares_best (mean (SD))              236112.80 (121726.01) 271367.86 (133997.25) <0.001       0.275
  kwshares_avg (mean (SD))                 2969.13 (1164.23)    3259.51 (1453.11)  <0.001       0.221
  self_reference_min_shares (mean (SD))    4027.48 (21038.58)   4022.10 (17739.38)  0.980      <0.001
  self_reference_max_shares (mean (SD))    8906.44 (35549.18)  12945.53 (49804.74) <0.001       0.093
  self_reference_avg_sharess (mean (SD))   5873.04 (23925.79)   7433.21 (25319.44) <0.001       0.063
  global_subjectivity (mean (SD))             0.44 (0.10)          0.44 (0.14)      0.018       0.024
  global_sentiment_polarity (mean (SD))       0.12 (0.09)          0.11 (0.10)     <0.001       0.109
  global_rate_positive_words (mean (SD))      0.04 (0.02)          0.04 (0.02)     <0.001       0.090
  global_rate_negative_words (mean (SD))      0.02 (0.01)          0.02 (0.01)     <0.001       0.088
  rate_positive_words (mean (SD))             0.70 (0.17)          0.65 (0.22)     <0.001       0.229
  rate_negative_words (mean (SD))             0.29 (0.15)          0.29 (0.17)      0.097       0.017
  avg_positive_polarity (mean (SD))           0.36 (0.09)          0.35 (0.12)     <0.001       0.049
  min_positive_polarity (mean (SD))           0.09 (0.07)          0.10 (0.08)     <0.001       0.043
  max_positive_polarity (mean (SD))           0.76 (0.23)          0.74 (0.28)     <0.001       0.071
  avg_negative_polarity (mean (SD))          -0.25 (0.12)         -0.27 (0.14)     <0.001       0.095
  min_negative_polarity (mean (SD))          -0.51 (0.28)         -0.53 (0.30)     <0.001       0.062
  max_negative_polarity (mean (SD))          -0.11 (0.09)         -0.11 (0.11)      0.003       0.031
  title_subjectivity (mean (SD))              0.27 (0.32)          0.31 (0.34)     <0.001       0.125
  title_sentiment_polarity (mean (SD))        0.07 (0.26)          0.07 (0.28)      0.760       0.003
  abs_title_subjectivity (mean (SD))          0.34 (0.19)          0.34 (0.19)      0.068       0.019
  abs_title_sentiment_polarity (mean (SD))    0.15 (0.22)          0.17 (0.24)     <0.001       0.094
  category (%)                                                                     <0.001       0.441
     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       0.077
     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)
```



Task 2 (a)

- From the table, we can see that the gap between the covariates of the two sample groups is not large, with the p-values generally being less than 0.001.
- From the overlap diagram, the common overlap between the two is concentrated in the range of 0.04 <= n <= 0.73, 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(0.04 <= n <= 0.73).
- In the result, there are 13,686 values corresponding to 0 and 13,686 values corresponding to 1, which is balanced.

```r
# Task 2b
# Perform matching by pscore
ds_matched <- ds[ds$pscore >= .04 & ds$pscore <= .73, ]
matched <- matchit(y_x, method = "nearest", data = ds)
ds_matched <- match.data(matched)
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 next page), I believe the propensity score matching process was successful because it effectively reduced the SMD, improved the balance between the two groups, and the corresponding p-value also increased. A reduction in SMD means that the differences in these covariates between the two groups have been effectively minimized, enhancing the accuracy of the estimation.

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

Task 2 (c)

```
> print(table_unmatched, smd = TRUE)
                                      Stratified by treat_ind
                                       0                    1                    p      test SMD
  n                                        25026                13686
  n_tokens_title (mean (SD))               10.21 (2.10)          10.69 (2.10)      <0.001        0.226
  n_tokens_content (mean (SD))            554.98 (449.36)       526.87 (501.64)    <0.001        0.059
  n_unique_tokens (mean (SD))               0.56 (4.43)           0.53 (0.17)       0.429        0.010
  n_non_stop_unique_tokens (mean (SD))      0.70 (4.11)           0.67 (0.19)       0.286        0.013
  num_hrefs (mean (SD))                    11.04 (11.31)         10.61 (11.34)     <0.001        0.038
  num_self_hrefs (mean (SD))                3.25 (4.06)           3.34 (3.52)       0.028        0.024
  num_imgs (mean (SD))                      4.68 (8.12)           4.20 (8.42)      <0.001        0.059
  average_token_length (mean (SD))          4.64 (0.63)           4.39 (1.11)      <0.001        0.283
  num_keywords (mean (SD))                  7.19 (1.94)           7.30 (1.84)      <0.001        0.057
  kwshares_worst (mean (SD))              318.09 (510.46)       302.34 (768.52)     0.016        0.024
  kwshares_best (mean (SD))            236112.80 (121726.01) 271367.86 (133997.25) <0.001        0.275
  kwshares_avg (mean (SD))               2969.13 (1164.23)     3259.51 (1453.11)   <0.001        0.221
  self_reference_min_shares (mean (SD))  4027.48 (21038.58)    4022.10 (17739.38)   0.980       <0.001
  self_reference_max_shares (mean (SD))  8906.44 (35549.18)   12945.53 (49804.74)  <0.001        0.093
  self_reference_avg_sharess (mean (SD)) 5873.04 (23925.79)    7433.21 (25319.44)  <0.001        0.063
  global_subjectivity (mean (SD))           0.44 (0.10)           0.44 (0.14)       0.018        0.024
  global_sentiment_polarity (mean (SD))     0.12 (0.09)           0.11 (0.10)      <0.001        0.109
  global_rate_positive_words (mean (SD))    0.04 (0.02)           0.04 (0.02)      <0.001        0.090
  global_rate_negative_words (mean (SD))    0.02 (0.01)           0.02 (0.01)      <0.001        0.088
  rate_positive_words (mean (SD))           0.70 (0.17)           0.65 (0.22)      <0.001        0.229
  rate_negative_words (mean (SD))           0.29 (0.15)           0.29 (0.17)       0.097        0.017
  avg_positive_polarity (mean (SD))         0.36 (0.09)           0.35 (0.12)      <0.001        0.049
  min_positive_polarity (mean (SD))         0.09 (0.07)           0.10 (0.08)      <0.001        0.043
  max_positive_polarity (mean (SD))         0.76 (0.23)           0.74 (0.28)      <0.001        0.071
  avg_negative_polarity (mean (SD))        -0.25 (0.12)          -0.27 (0.14)      <0.001        0.095
  min_negative_polarity (mean (SD))        -0.51 (0.28)          -0.53 (0.30)      <0.001        0.062
  max_negative_polarity (mean (SD))        -0.11 (0.09)          -0.11 (0.11)       0.003        0.031
  title_subjectivity (mean (SD))            0.27 (0.32)           0.31 (0.34)      <0.001        0.125
  title_sentiment_polarity (mean (SD))      0.07 (0.26)           0.07 (0.28)       0.760        0.003
  abs_title_subjectivity (mean (SD))        0.34 (0.19)           0.34 (0.19)       0.068        0.019
  abs_title_sentiment_polarity (mean (SD))  0.15 (0.22)           0.17 (0.24)      <0.001        0.094
  category (%)                                                                     <0.001        0.441
     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        0.077
     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)
```

```
> print(table_matched, smd = TRUE)
                                      Stratified by treat_ind
                                       0                    1                    p      test SMD
  n                                        13686                13686
  n_tokens_title (mean (SD))               10.55 (2.11)          10.69 (2.10)      <0.001        0.064
  n_tokens_content (mean (SD))            553.93 (457.12)       526.87 (501.64)    <0.001        0.056
  n_unique_tokens (mean (SD))               0.58 (5.99)           0.53 (0.17)       0.356        0.011
  n_non_stop_unique_tokens (mean (SD))      0.72 (5.55)           0.67 (0.19)       0.288        0.013
  num_hrefs (mean (SD))                    11.02 (11.40)         10.61 (11.34)      0.003        0.036
  num_self_hrefs (mean (SD))                3.40 (4.45)           3.34 (3.52)       0.263        0.014
  num_imgs (mean (SD))                      4.57 (7.19)           4.20 (8.42)      <0.001        0.047
  average_token_length (mean (SD))          4.55 (0.79)           4.39 (1.11)      <0.001        0.169
  num_keywords (mean (SD))                  7.28 (1.96)           7.30 (1.84)       0.602        0.006
  kwshares_worst (mean (SD))              300.11 (591.01)       302.34 (768.52)     0.787        0.003
  kwshares_best (mean (SD))            254273.47 (117490.41) 271367.86 (133997.25) <0.001        0.136
  kwshares_avg (mean (SD))               3107.92 (1276.15)     3259.51 (1453.11)   <0.001        0.111
  self_reference_min_shares (mean (SD))  4214.66 (21212.23)    4022.10 (17739.38)   0.415        0.010
  self_reference_max_shares (mean (SD)) 10396.47 (43135.68)   12945.53 (49804.74)  <0.001        0.055
  self_reference_avg_sharess (mean (SD)) 6546.76 (25996.14)    7433.21 (25319.44)   0.004        0.035
  global_subjectivity (mean (SD))           0.45 (0.11)           0.44 (0.14)      <0.001        0.070
  global_sentiment_polarity (mean (SD))     0.12 (0.10)           0.11 (0.10)       0.002        0.038
  global_rate_positive_words (mean (SD))    0.04 (0.02)           0.04 (0.02)      <0.001        0.045
  global_rate_negative_words (mean (SD))    0.02 (0.01)           0.02 (0.01)       0.968       <0.001
  rate_positive_words (mean (SD))           0.68 (0.18)           0.65 (0.22)      <0.001        0.113
  rate_negative_words (mean (SD))           0.30 (0.15)           0.29 (0.17)      <0.001        0.049
  avg_positive_polarity (mean (SD))         0.36 (0.10)           0.35 (0.12)      <0.001        0.063
  min_positive_polarity (mean (SD))         0.10 (0.07)           0.10 (0.08)       0.632        0.006
  max_positive_polarity (mean (SD))         0.76 (0.24)           0.74 (0.28)      <0.001        0.068
  avg_negative_polarity (mean (SD))        -0.27 (0.13)          -0.27 (0.14)       0.124        0.019
  min_negative_polarity (mean (SD))        -0.55 (0.29)          -0.53 (0.30)      <0.001        0.043
  max_negative_polarity (mean (SD))        -0.11 (0.10)          -0.11 (0.11)       0.998       <0.001
  title_subjectivity (mean (SD))            0.30 (0.33)           0.31 (0.34)       0.004        0.035
  title_sentiment_polarity (mean (SD))      0.07 (0.27)           0.07 (0.28)       0.296        0.013
  abs_title_subjectivity (mean (SD))        0.34 (0.19)           0.34 (0.19)       0.631        0.006
  abs_title_sentiment_polarity (mean (SD))  0.16 (0.23)           0.17 (0.24)       0.004        0.034
  category (%)                                                                     <0.001        0.063
     business                               1403 (10.3)           1365 (10.0)
     entertainment                          2751 (20.1)           2973 (21.7)
     lifestyle                               515 ( 3.8)            468 ( 3.4)
     socialmedia                             754 ( 5.5)            681 ( 5.0)
     tech                                   2280 (16.7)           2071 (15.1)
     world                                  5983 (43.7)           6128 (44.8)
  weekday (%)                                                                       0.177        0.036
     friday                                 1996 (14.6)           2008 (14.7)
     monday                                 2258 (16.5)           2313 (16.9)
     saturday                                796 ( 5.8)            724 ( 5.3)
     sunday                                  933 ( 6.8)            853 ( 6.2)
     thursday                               2512 (18.4)           2505 (18.3)
     tuesday                                2603 (19.0)           2683 (19.6)
     wednesday                              2588 (18.9)           2600 (19.0)
```

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.

```r
# 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 1320.8 shares.

```
> summary(model_t3)

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

Residuals:
   Min      1Q Median      3Q    Max
 -4309   -2810  -1989    -689 838990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2989.3      114.3   26.15  < 2e-16 ***
treat_ind     1320.8      161.7    8.17  3.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13370 on 27370 degrees of freedom
Multiple R-squared:  0.002433,  Adjusted R-squared:  0.002397
F-statistic: 66.76 on 1 and 27370 DF,  p-value: 3.205e-16
```

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. Due to copyright and other reasons, some articles cannot legally cite videos. Some video materials may have strict copyright restrictions; therefore, these articles can only display images or textual descriptions, which results in the treatment indicator being 0.
  2. Some articles do not require the use of videos because text and images are sufficient to outline the content for specific readers, which also leads to the absence of videos in the articles.
- All the above are random factors that could affect the treatment indicator but are not included in the dataset. Moreover, they are not directly related to shares.

Thank you

Beiming Zhang

BU.450.760.K1.SP24