TSTV-Obs-Feb_26_DM.R

danny

2020-02-26

```r
#*** MSBA 6440 ***#
#*** Gordon Burtch and Gautam Ray***#
#*** Updated Feb 2020 ***#
#*** Code for Lecture 4 ***#
#*** Propensity Score Matching ***#


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
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

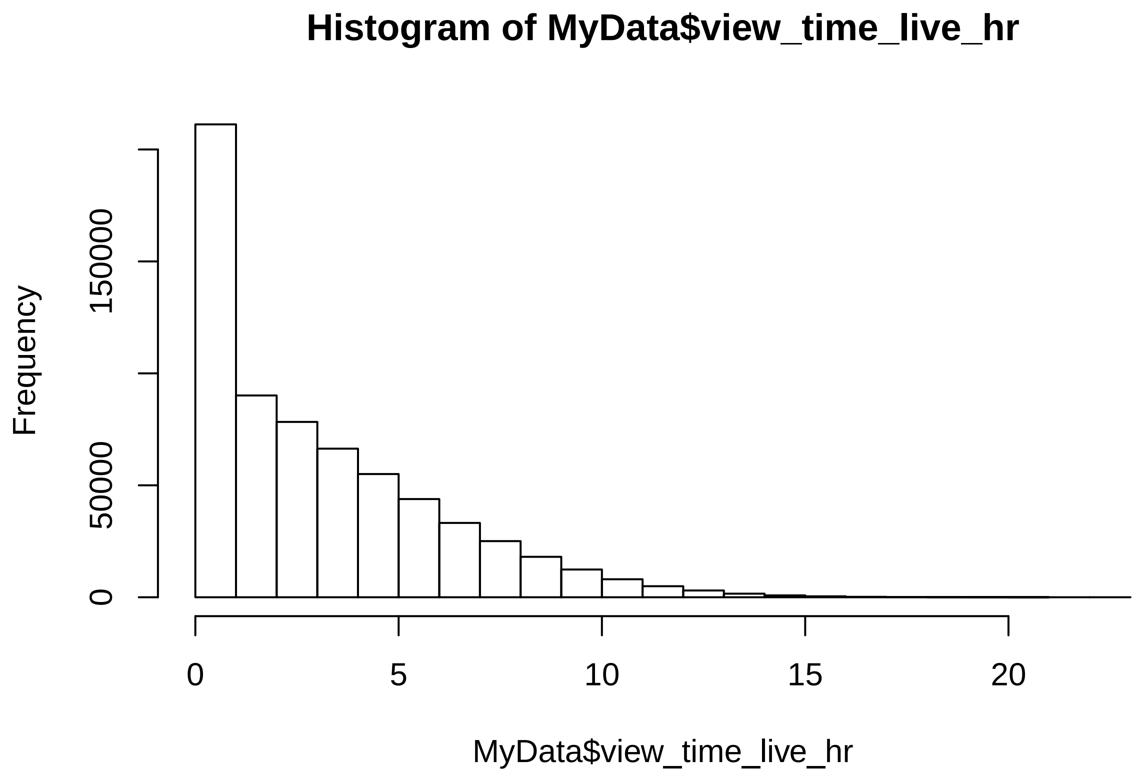
```r
library(MatchIt)
```

```
## Warning: package 'MatchIt' was built under R version 3.6.2
```

```r
library(data.table)
library(tableone)
```

```
## Warning: package 'tableone' was built under R version 3.6.2
```

```r
#**** Load the data ***#
MyData<-read.csv("TSTV-Obs-Dataset-2.csv")

hist(MyData$view_time_live_hr)
```




```r
#*** Let's get a sense of the data ***#

#how long is the period of observation?
max(MyData$week)-min(MyData$week)
```

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

```r
#How many subjects got TSTV? (Treated)
length(unique(MyData$id[MyData$premium==TRUE]))
```

```
## [1] 8348
```

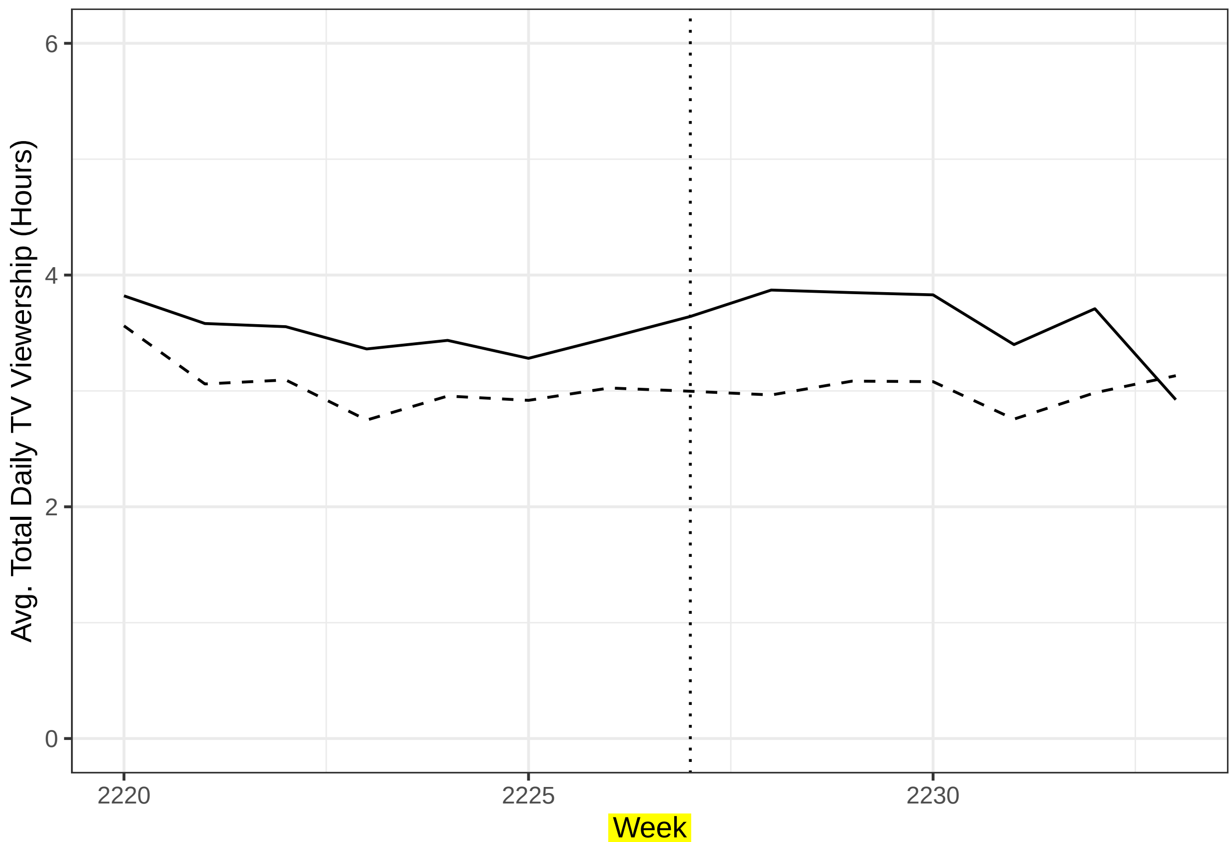
```r
#How many subjects did not get TSTV? (Control)
length(unique(MyData$id[MyData$premium==FALSE]))
```

```
## [1] 41686
```

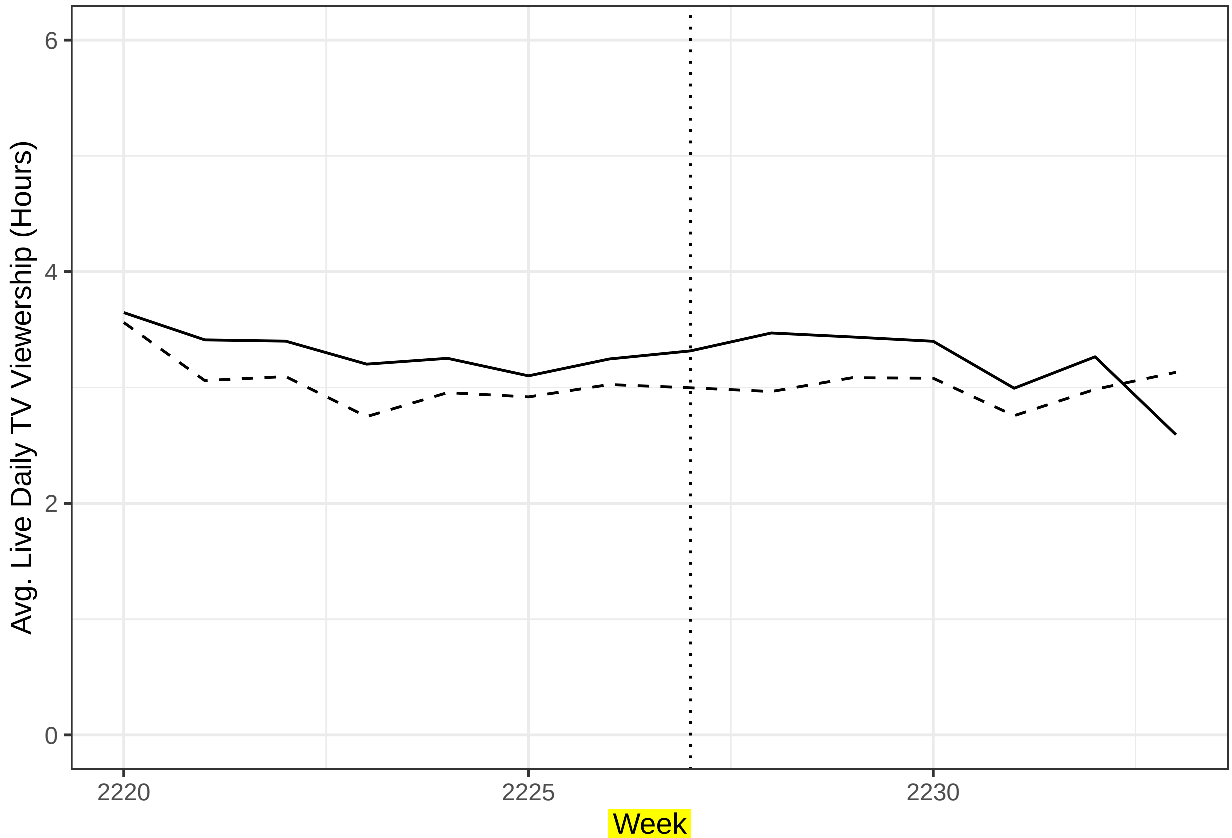
```r
#In what 'week' does the "treatment" begin?
min(unique(MyData$week[MyData$after==TRUE]))
```

```
## [1] 2227
```

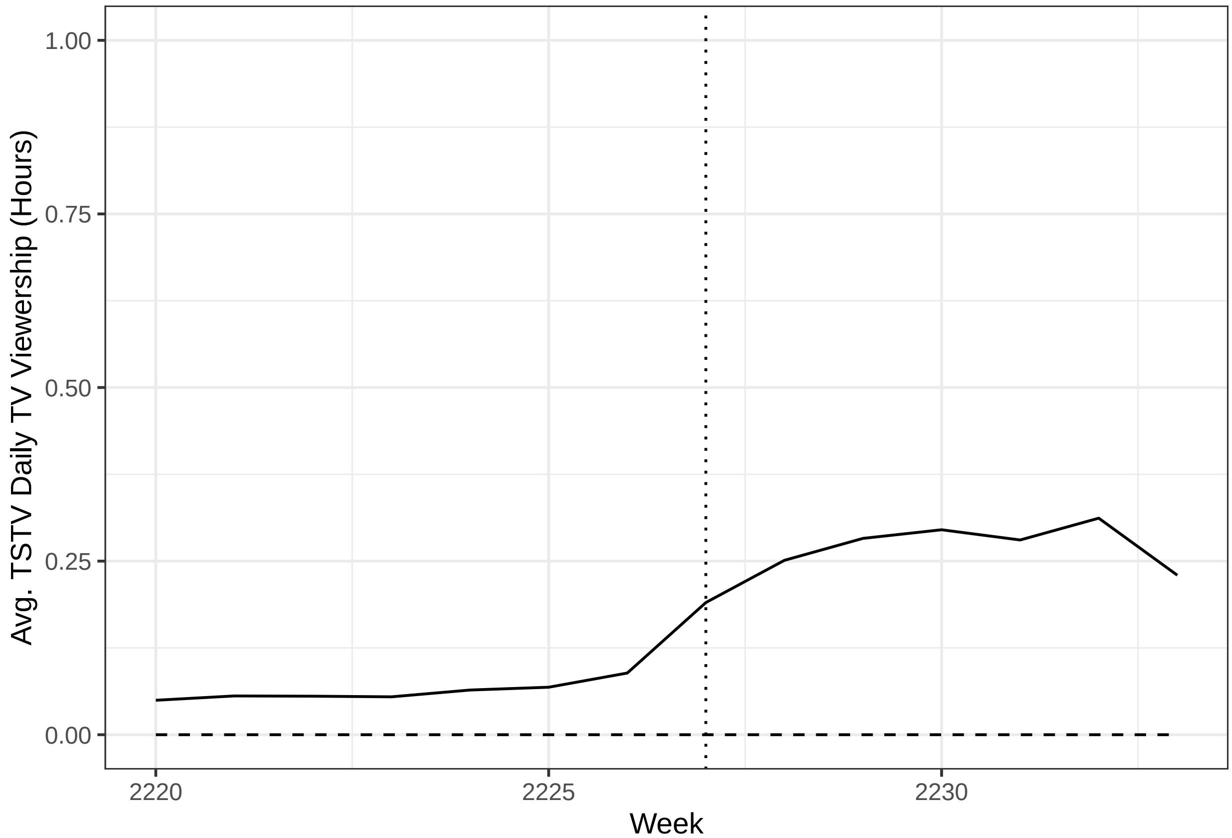
```r
#Let's just look at what is going on with average viewership behavior
#for treated vs. untreated, in the weeks around the treatment date.
MyDataAggregated <- aggregate(MyData,by=list(MyData$premium, MyData$week),FUN=mean)

# plot for total TV time
p <- ggplot(MyDataAggregated)
p <- p + geom_line(data=MyDataAggregated[MyDataAggregated$premium==FALSE,], aes(week, view_time_total_hr
p <- p + geom_line(data=MyDataAggregated[MyDataAggregated$premium==TRUE,], aes(week, view_time_total_hr)
p <- p + geom_vline(xintercept=2227, linetype='dotted')
p <- p  + xlab("Week") + ylab("Avg. Total Daily TV Viewership (Hours)")
p <- p + ylim(0, 6) + xlim(2220,2233) + theme_bw()
p
```




```r
# plot for live TV time
p <- ggplot(MyDataAggregated)
p <- p + geom_line(data=MyDataAggregated[MyDataAggregated$premium==FALSE,], aes(week, view_time_live_hr)
p <- p + geom_line(data=MyDataAggregated[MyDataAggregated$premium==TRUE,], aes(week, view_time_live_hr)
p <- p + geom_vline(xintercept=2227, linetype='dotted')
p <- p  + xlab("Week") + ylab("Avg. Live Daily TV Viewership (Hours)")
p <- p + ylim(0, 6) + xlim(2220,2233) + theme_bw()
p
```




```r
# plot for TSTV time
p <- ggplot(MyDataAggregated)
p <- p + geom_line(data=MyDataAggregated[MyDataAggregated$premium==FALSE,], aes(week, view_time_tstv_hr
p <- p + geom_line(data=MyDataAggregated[MyDataAggregated$premium==TRUE,], aes(week, view_time_tstv_hr)
p <- p + geom_vline(xintercept=2227, linetype='dotted')
p <- p  + xlab("Week") + ylab("Avg. TSTV Daily TV Viewership (Hours)")
p <- p + ylim(0, 1) + xlim(2220,2233) + theme_bw()
p
```



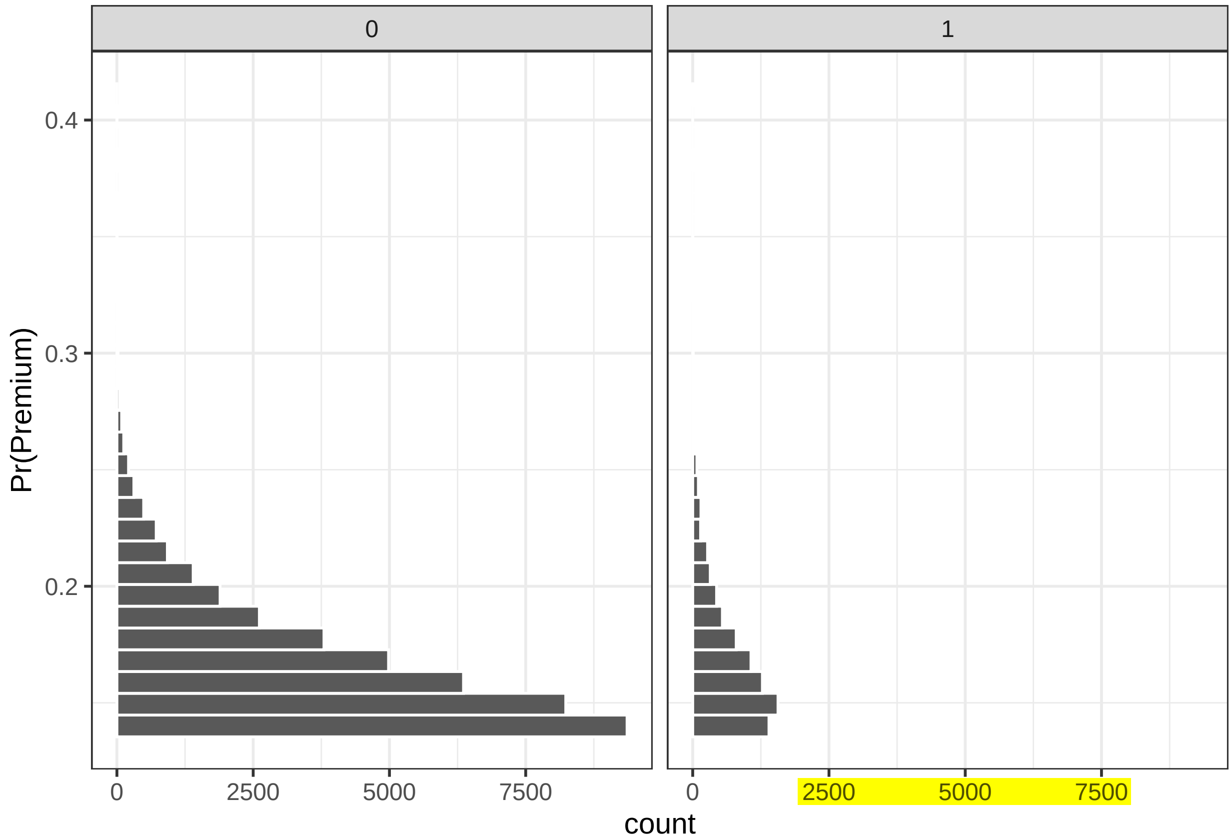

```
#*** Propensity Score Matching ***#

#For this demonstration, we will use data from the pre-period for matching.
#We will then estimate the effect of TSTV gifting in the post period.

#*** CREATE A SUMMARY DATASET BEFORE vs. AFTER TSTV IS AVAILABLE ***#
MyDataSummary <- aggregate(MyData,by=list(MyData$id,MyData$after),FUN=mean)
MyDataSummary$view_time_total_sq <- MyDataSummary$view_time_total_hr^2


# Okay, let's check out our covariate balance; we have one confounder here, view_time_total_hr.
# This is a dependent variable, but we are going to match on it in the pre-period.
# That is, we only want subjects who had similar viewership activity before TSTV showed up.

MyPreData <- MyDataSummary[MyDataSummary$after == FALSE,]

tabUnmatched <- CreateTableOne(vars=c("view_time_total_hr","view_time_total_sq"), strata="premium", tes
print(tabUnmatched, smd=TRUE)
```

```
##                                 Stratified by premium
##                                  0             1             p      test
##   n                              41686          8348
##   view_time_total_hr (mean (SD))  2.98 (2.42)   3.46 (2.69)  <0.001
##   view_time_total_sq (mean (SD)) 14.72 (22.20) 19.21 (28.21) <0.001
##                                 Stratified by premium
##                                  SMD
```

```
##   n
##   view_time_total_hr (mean (SD))  0.191
##   view_time_total_sq (mean (SD))  0.177
```

```
# Whoa, lots of imbalance here...

# Let's see what propensity scores look like...


MyPreData$PS<-glm(premium~view_time_total_hr+view_time_total_sq, data=MyPreData, family = "binomial")$fi
ggplot(MyPreData, aes(x = PS)) +
  geom_histogram(color = "white") +
  facet_wrap(~premium) + xlab("Pr(Premium)") +theme_bw() + coord_flip()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



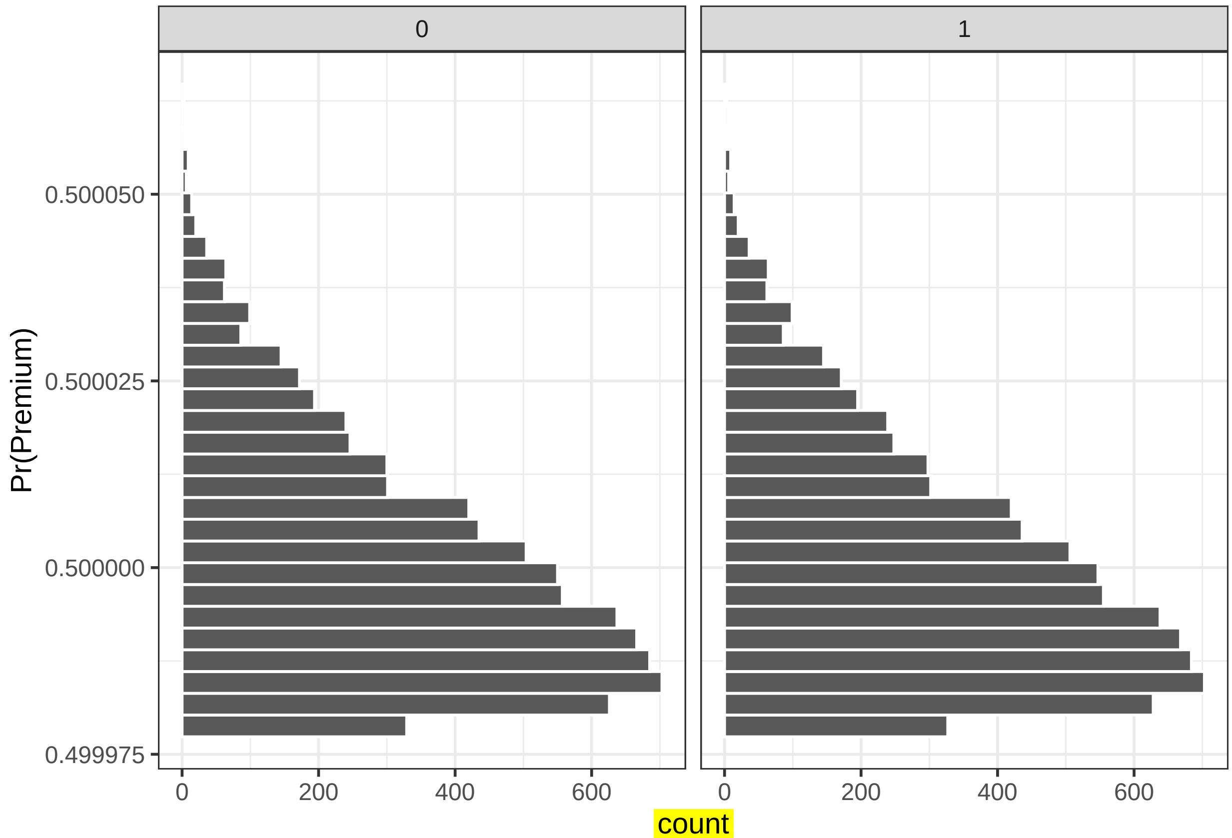

```
#*** Match treated and control households on propensity to receive premium based on pre-treatment time s
# Note: the matchit command may take a long time to run with large datasets

Matched_Output <- matchit(premium ~ view_time_total_hr + view_time_total_sq, data = MyPreData, method =
summary(Matched_Output)
```

```
##
## Call:
```

```
## matchit(formula = premium ~ view_time_total_hr + view_time_total_sq,
##     data = MyPreData, method = "nearest", distance = "logit",
##     caliper = 0.001, replace = FALSE)
##
## Summary of balance for all data:
##                    Means Treated Means Control SD Control Mean Diff
## distance                  0.1714        0.1659     0.0267    0.0055
## view_time_total_hr        3.4632        2.9754     2.4220    0.4878
## view_time_total_sq       19.2077       14.7191    22.1998    4.4887
##                    eQQ Med eQQ Mean eQQ Max
## distance            0.0042   0.0055  0.0288
## view_time_total_hr  0.4336   0.4873  1.6377
## view_time_total_sq  2.3247   4.4755 39.9870
##
##
## Summary of balance for matched data:
##                    Means Treated Means Control SD Control Mean Diff
## distance                  0.1686        0.1686     0.0262    0.0000
## view_time_total_hr        3.2484        3.2483     2.3880    0.0001
## view_time_total_sq       16.2544       16.2533    21.4272    0.0011
##                    eQQ Med eQQ Mean eQQ Max
## distance            0.0000   0.0000  0.0000
## view_time_total_hr  0.0007   0.0008  0.0031
## view_time_total_sq  0.0031   0.0055  0.0369
##
## Percent Balance Improvement:
##                    Mean Diff. eQQ Med eQQ Mean eQQ Max
## distance              99.9718 99.8336  99.8466 99.9042
## view_time_total_hr    99.9699 99.8347  99.8336 99.8121
## view_time_total_sq    99.9762 99.8665  99.8762 99.9076
##
## Sample sizes:
##           Control Treated
## All         41686    8348
## Matched      8110    8110
## Unmatched   33576     238
## Discarded       0       0
```

```r
Matched.ids <- data.table(match.data(Matched_Output))$id

Matched_Data = match.data(Matched_Output)

Matched_Data$PS = glm(premium ~ view_time_total_hr, data = Matched_Data, family = "binomial")$fitted.val
ggplot(Matched_Data, aes(x = PS)) +
  geom_histogram(color = "white") +
  facet_wrap(~premium) + xlab("Pr(Premium)") +theme_bw() + coord_flip()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
tabMatched <- CreateTableOne(vars=c("view_time_total_hr","view_time_total_sq"), strata="premium", test=
print(tabMatched, smd=TRUE)
```

```
##                                   Stratified by premium
##                                    0             1             p      test
##   n                                 8110          8110
##   view_time_total_hr (mean (SD))  3.25 (2.39)   3.25 (2.39)   0.997
##   view_time_total_sq (mean (SD)) 16.25 (21.43) 16.25 (21.43)  0.997
##                                   Stratified by premium
##                                    SMD
##   n
##   view_time_total_hr (mean (SD)) <0.001
##   view_time_total_sq (mean (SD)) <0.001
```

```
#Now let's estimate the treatment effect with vs. without matching.
MyDataPost <- MyDataSummary[MyDataSummary$after==TRUE,]
unmatched_ate <- lm(data=MyDataPost,view_time_total_hr~premium)
matched_ate <- lm(data=MyDataPost[MyDataPost$id %in% Matched.ids,], view_time_total_hr ~ premium)

#Produce the output table.
stargazer(unmatched_ate,matched_ate,title="Matched vs. Unmatched Estimates",column.labels=c("Total View
```

```
##
## Matched vs. Unmatched Estimates
## ========================================================================
```

```
##                                     Dependent variable:
##                       ----------------------------------------------------
##                                       view_time_total_hr
##                            Total Viewership          Total Viewership
##                                  (1)                       (2)
## ---------------------------------------------------------------------------
## premium                        0.614***                  0.199***
##                                (0.031)                   (0.040)
##
## Constant                       2.990***                  3.238***
##                                (0.013)                   (0.029)
##
## ---------------------------------------------------------------------------
## Observations                    48,483                    15,914
## R2                              0.008                     0.002
## Adjusted R2                     0.008                     0.001
## Residual Std. Error     2.586 (df = 48481)        2.549 (df = 15912)
## F Statistic          388.341*** (df = 1; 48481) 24.130*** (df = 1; 15912)
## ===========================================================================
## Note:                                         *p<0.1; **p<0.05; ***p<0.01
```