highnote.R

DoryChen
2019-05-10

```
setwd("/Users/DoryChen/Desktop")

data_read <- read.csv("HighNote Data Midterm.csv", header=T)

#install packages

library(MatchIt)
library(pastecs)
library(dplyr)
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:pastecs':
##
##     first, last
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
library(ggplot2)
library(rmarkdown)
## Warning: package 'rmarkdown' was built under R version 3.5.2
#Pre-analysis
data_read$subscriber_friend <- ifelse(data_read$subscriber_friend_cnt == 0, 0,1)
table(data_read$subscriber_friend)
##
##     0     1
## 34004  9823
t.test(data_read$adopter~data_read$subscriber_friend)
##
##  Welch Two Sample t-test
##
## data:  data_read$adopter by data_read$subscriber_friend
## t = -30.961, df = 11815, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1330281 -0.1171869
## sample estimates:
## mean in group 0 mean in group 1
##      0.05243501      0.17754250
adopter_cov <- c('age', 'male', 'friend_cnt', 'avg_friend_age','avg_friend_male',
'friend_country_cnt',"subscriber_friend_cnt","songsListened",'lovedTracks',"posts".'playlists',"shou
ts",'tenure','good_country')
lapply(adopter_cov, function(v) {
 t.test(data_read[, v] ~ data_read$adopter)
})
## [[1]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -16.996, df = 4079.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.265768 -1.797097
## sample estimates:
## mean in group 0 mean in group 1
##        23.94844        25.97987
##
##
## [[2]]
##
##  Welch Two Sample t-test
##
```

```
## data:  data_read[, v] by data_read$adopter
## t = -13.654, df = 4295, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.12278707 -0.09195413
## sample estimates:
## mean in group 0 mean in group 1
##       0.6218610       0.7292316
##
##
## [[3]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -10.646, df = 3675.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -25.15422 -17.32999
## sample estimates:
## mean in group 0 mean in group 1
##        18.49166        39.73377
##
##
## [[4]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -15.658, df = 4140.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.608931 -1.250852
## sample estimates:
## mean in group 0 mean in group 1
##        24.01142        25.44131
##
##
## [[5]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -4.4426, df = 4591.6, p-value = 9.097e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02883955 -0.01117951
## sample estimates:
## mean in group 0 mean in group 1
##       0.6165888       0.6365983
##
##
## [[6]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -21.267, df = 3791.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.528795 -2.933081
## sample estimates:
## mean in group 0 mean in group 1
##        3.957891        7.188829
##
##
## [[7]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -12.287, df = 3632.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.413899 -1.024766
```

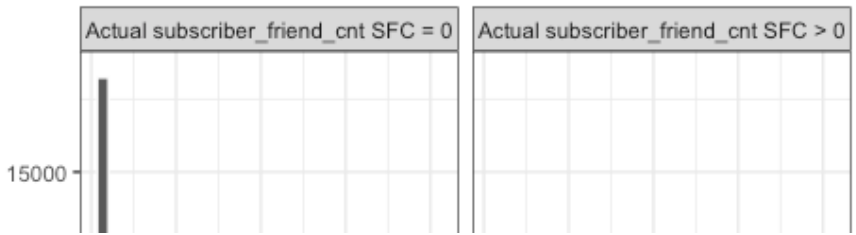
```
## sample estimates:
## mean in group 0 mean in group 1
##        0.417469        1.636802
##
##
## [[8]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -21.629, df = 3792.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -17634.24 -14702.96
## sample estimates:
## mean in group 0 mean in group 1
##        17589.44        33758.04
##
##
## [[9]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -21.188, df = 3705.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -193.9447 -161.0917
## sample estimates:
## mean in group 0 mean in group 1
##        86.82263       264.34080
##
##
## [[10]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -4.2151, df = 3663.5, p-value = 2.557e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -23.30665  -8.50825
## sample estimates:
## mean in group 0 mean in group 1
##        5.293002       21.200454
##
##
## [[11]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -8.0816, df = 3634.7, p-value = 8.619e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4367565 -0.2662138
## sample estimates:
## mean in group 0 mean in group 1
##       0.5492804       0.9007655
##
##
## [[12]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -3.5659, df = 3536.5, p-value = 0.0003674
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -107.66170  -31.27249
## sample estimates:
## mean in group 0 mean in group 1
##        29.97266        99.43975
##
##
## [[13]]
```

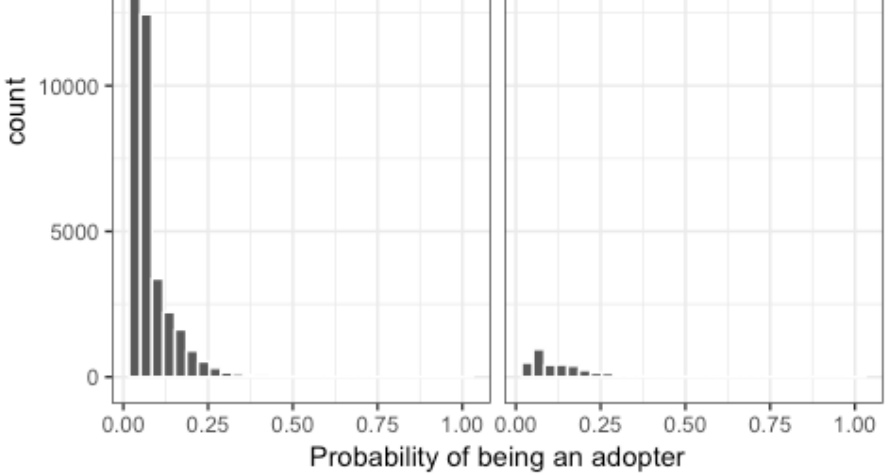
```
## [[13]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = -5.0434, df = 4150.6, p-value = 4.768e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.462620 -1.083959
## sample estimates:
## mean in group 0 mean in group 1
##        43.80993        45.58322
##
##
## [[14]]
##
##  Welch Two Sample t-test
##
## data:  data_read[, v] by data_read$adopter
## t = 8.8009, df = 4248.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05463587 0.08595434
## sample estimates:
## mean in group 0 mean in group 1
##       0.3577916       0.2874965
# statistic descriptive

adopt<-filter(data_read,adopter==1)
non_adopter<-filter(data_read,adopter==0)

summary_adopt<-stat.desc(adopt)
summary_non_adopt<-stat.desc(non_adopter)

summary_adopt
##                      ID         age        male   friend_cnt
## nbr.val      3.527000e+03 3.527000e+03 3.527000e+03 3.527000e+03
## nbr.null     0.000000e+00 0.000000e+00 9.550000e+02 0.000000e+00
## nbr.na       0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## min          4.030100e+04 8.000000e+00 0.000000e+00 1.000000e+00
## max          4.382700e+04 7.300000e+01 1.000000e+00 5.089000e+03
## range        3.526000e+03 6.500000e+01 1.000000e+00 5.088000e+03
## sum          1.483597e+08 9.163100e+04 2.572000e+03 1.401410e+05
## median       4.206400e+04 2.400000e+01 1.000000e+00 1.600000e+01
## mean         4.206400e+04 2.597987e+01 7.292316e-01 3.973377e+01
## SE.mean      1.714643e+01 1.152343e-01 7.483255e-03 1.974705e+00
## CI.mean.0.95 3.361792e+01 2.259326e-01 1.467195e-02 3.871681e+00
## var          1.036938e+06 4.683482e+01 1.975089e-01 1.375340e+04
## std.dev      1.018302e+03 6.843597e+00 4.444197e-01 1.172749e+02
## coef.var     2.420839e-02 2.634192e-01 6.094355e-01 2.951517e+00
##              avg_friend_age avg_friend_male friend_country_cnt
## nbr.val       3.527000e+03    3.527000e+03       3.527000e+03
## nbr.null      0.000000e+00    1.300000e+02       7.000000e+00
## nbr.na        0.000000e+00    0.000000e+00       0.000000e+00
## min           1.200000e+01    0.000000e+00       0.000000e+00
## max           6.200000e+01    1.000000e+00       1.360000e+02
## range         5.000000e+01    1.000000e+00       1.360000e+02
## sum           8.973150e+04    2.245282e+03       2.535500e+04
## median        2.436000e+01    6.666667e-01       4.000000e+00
## mean          2.544131e+01    6.365983e-01       7.188829e+00
## SE.mean       8.771087e-02    4.214407e-03       1.491839e-01
## CI.mean.0.95  1.719692e-01    8.262923e-03       2.924956e-01
## var           2.713390e+01    6.264385e-02       7.849638e+01
## std.dev       5.209021e+00    2.502875e-01       8.859818e+00
## coef.var      2.047466e-01    3.931640e-01       1.232442e+00
##              subscriber_friend_cnt songsListened  lovedTracks        posts
## nbr.val               3.527000e+03  3.527000e+03 3.527000e+03  3527.000000
## nbr.null              1.783000e+03  1.000000e+00 1.970000e+02  2158.000000
## nbr.na                0.000000e+00  0.000000e+00 0.000000e+00     0.000000
## min                   0.000000e+00  0.000000e+00 0.000000e+00     0.000000
## max                   2.870000e+02  8.172900e+05 1.022000e+04  8506.000000
## range                 2.870000e+02  8.172900e+05 1.022000e+04  8506.000000
## sum                   5.773000e+03  1.190646e+08 9.323300e+05 74774.000000
## median                0.000000e+00  2.090800e+04 1.080000e+02     0.000000
## mean                  1.636802e+00  3.375804e+04 2.643408e+02    21.200454
## SE.mean               9.850351e-02  7.340258e+02 8.274773e+00     3.737984
```

```
## CI.mean.0.95          1.931296e-01  1.439158e+03 1.622383e+01     7.328829
## var                   3.422228e+01  1.900326e+09 2.415003e+05 49281.089416
## std.dev               5.849981e+00  4.359273e+04 4.914268e+02   221.993445
## coef.var              3.574031e+00  1.291329e+00 1.859065e+00    10.471165
##                playlists       shouts adopter       tenure good_country
## nbr.val     3.527000e+03 3.527000e+03    3527 3.527000e+03 3.527000e+03
## nbr.null    1.598000e+03 2.410000e+02       0 1.000000e+00 2.513000e+03
## nbr.na      0.000000e+00 0.000000e+00       0 0.000000e+00 0.000000e+00
## min         0.000000e+00 0.000000e+00       1 0.000000e+00 0.000000e+00
## max         1.180000e+02 6.587200e+04       1 1.110000e+02 1.000000e+00
## range       1.180000e+02 6.587200e+04       0 1.110000e+02 1.000000e+00
## sum         3.177000e+03 3.507240e+05    3527 1.607720e+05 1.014000e+03
## median      1.000000e+00 9.000000e+00       1 4.600000e+01 0.000000e+00
## mean        9.007655e-01 9.943975e+01       1 4.558322e+01 2.874965e-01
## SE.mean     4.316306e-02 1.946626e+01       0 3.375022e-01 7.621994e-03
## CI.mean.0.95 8.462710e-02 3.816627e+01       0 6.617192e-01 1.494396e-02
## var         6.570978e+00 1.336505e+06       0 4.017525e+02 2.049003e-01
## std.dev     2.563392e+00 1.156073e+03       0 2.004376e+01 4.526592e-01
## coef.var    2.845793e+00 1.162587e+01       0 4.397181e-01 1.574486e+00
##             subscriber_friend
## nbr.val          3.527000e+03
## nbr.null         1.783000e+03
## nbr.na           0.000000e+00
## min              0.000000e+00
## max              1.000000e+00
## range            1.000000e+00
## sum              1.744000e+03
## median           0.000000e+00
## mean             4.944712e-01
## SE.mean          8.419810e-03
## CI.mean.0.95     1.650819e-02
## var              2.500403e-01
## std.dev          5.000403e-01
## coef.var         1.011263e+00
summary_non_adopt
##                   ID         age         male   friend_cnt
## nbr.val     4.030000e+04 4.030000e+04 4.030000e+04 4.030000e+04
## nbr.null    0.000000e+00 0.000000e+00 1.523900e+04 0.000000e+00
## nbr.na      0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## min         1.000000e+00 8.000000e+00 0.000000e+00 1.000000e+00
## max         4.030000e+04 7.900000e+01 1.000000e+00 4.957000e+03
## range       4.029900e+04 7.100000e+01 1.000000e+00 4.956000e+03
## sum         8.120652e+08 9.651220e+05 2.506100e+04 7.452140e+05
## median      2.015050e+04 2.300000e+01 1.000000e+00 7.000000e+00
## mean        2.015050e+04 2.394844e+01 6.218610e-01 1.849166e+01
## SE.mean     5.795185e+01 3.174035e-02 2.415601e-03 2.863341e-01
## CI.mean.0.95 1.135869e+02 6.221182e-02 4.734634e-03 5.612214e-01
## var         1.353442e+08 4.060023e+01 2.351557e-01 3.304085e+03
## std.dev     1.163375e+04 6.371831e+00 4.849286e-01 5.748117e+01
## coef.var    5.773431e-01 2.660646e-01 7.798021e-01 3.108491e+00
##             avg_friend_age avg_friend_male friend_country_cnt
## nbr.val        4.030000e+04    4.030000e+04       4.030000e+04
## nbr.null       0.000000e+00    4.398000e+03       2.620000e+02
## nbr.na         0.000000e+00    0.000000e+00       0.000000e+00
## min            8.000000e+00    0.000000e+00       0.000000e+00
## max            7.700000e+01    1.000000e+00       1.290000e+02
## range          6.900000e+01    1.000000e+00       1.290000e+02
## sum            9.676601e+05    2.484853e+04       1.595030e+05
## median         2.300000e+01    6.666667e-01       2.000000e+00
## mean           2.401142e+01    6.165888e-01       3.957891e+00
## SE.mean        2.542538e-02    1.588977e-03       2.871336e-02
## CI.mean.0.95   4.983432e-02    3.114432e-03       5.627885e-02
## var            2.605192e+01    1.017514e-01       3.322563e+01
## std.dev        5.104109e+00    3.189849e-01       5.764167e+00
## coef.var       2.125701e-01    5.173382e-01       1.456374e+00
##             subscriber_friend_cnt songsListened  lovedTracks        posts
## nbr.val              4.030000e+04  4.030000e+04 4.030000e+04 4.030000e+04
## nbr.null             3.222100e+04  1.446000e+03 9.607000e+03 3.146400e+04
## nbr.na               0.000000e+00  0.000000e+00 0.000000e+00 0.000000e+00
## min                  0.000000e+00  0.000000e+00 0.000000e+00 0.000000e+00
## max                  3.090000e+02  1.000000e+06 1.252200e+04 1.230900e+04
## range                3.090000e+02  1.000000e+06 1.252200e+04 1.230900e+04
## sum                  1.682400e+04  7.088545e+08 3.498952e+06 2.133080e+05
## median               0.000000e+00  7.440000e+03 1.400000e+01 0.000000e+00
## mean                 4.174690e-01  1.758944e+04 8.682263e+01 5.293002e+00
```

```
## SE.mean              1.204567e-02  1.415503e+02 1.312988e+00 5.196023e-01
## CI.mean.0.95         2.360978e-02  2.774418e+02 2.573486e+00 1.018432e+00
## var                  5.847453e+00  8.074704e+08 6.947465e+04 1.088046e+04
## std.dev              2.418151e+00  2.841602e+04 2.635804e+02 1.043094e+02
## coef.var             5.792408e+00  1.615516e+00 3.035850e+00 1.970704e+01
##                playlists       shouts adopter       tenure good_country
## nbr.val     4.030000e+04 4.030000e+04   40300 4.030000e+04 4.030000e+04
## nbr.null    2.188000e+04 3.311000e+03   40300 0.000000e+00 2.588100e+04
## nbr.na      0.000000e+00 0.000000e+00       0 0.000000e+00 0.000000e+00
## min         0.000000e+00 0.000000e+00       0 1.000000e+00 0.000000e+00
## max         9.800000e+01 7.736000e+03       0 1.110000e+02 1.000000e+00
## range       9.800000e+01 7.736000e+03       0 1.100000e+02 1.000000e+00
## sum         2.213600e+04 1.207898e+06       0 1.765540e+06 1.441900e+04
## median      0.000000e+00 4.000000e+00       0 4.400000e+01 0.000000e+00
## mean        5.492804e-01 2.997266e+01       0 4.380993e+01 3.577916e-01
## SE.mean     5.339791e-03 7.506393e-01       0 9.857536e-02 2.387844e-03
## CI.mean.0.95 1.046611e-02 1.471270e+00       0 1.932100e-01 4.680228e-03
## var         1.149089e+00 2.270741e+04       0 3.915992e+02 2.297825e-01
## std.dev     1.071956e+00 1.506898e+02       0 1.978887e+01 4.793563e-01
## coef.var    1.951564e+00 5.027576e+00     NaN 4.516982e-01 1.339764e+00
##             subscriber_friend
## nbr.val          4.030000e+04
## nbr.null         3.222100e+04
## nbr.na           0.000000e+00
## min              0.000000e+00
## max              1.000000e+00
## range            1.000000e+00
## sum              8.079000e+03
## median           0.000000e+00
## mean             2.004715e-01
## SE.mean          1.994326e-03
## CI.mean.0.95     3.908924e-03
## var              1.602866e-01
## std.dev          4.003581e-01
## coef.var         1.997083e+00
# Propensity Score Estimation
# Estimate the propensity score by running a logistic model

mylogit<-glm(adopter~ age + male + friend_cnt + avg_friend_age + avg_friend_male +
subscriber_friend + friend_country_cnt + songsListened
                + lovedTracks + posts + playlists + shouts + tenure +
good_country,data=data_read,family=binomial())


summary(mylogit)
##
## Call:
## glm(formula = adopter ~ age + male + friend_cnt + avg_friend_age +
##     avg_friend_male + subscriber_friend + friend_country_cnt +
##     songsListened + lovedTracks + posts + playlists + shouts +
##     tenure + good_country, family = binomial(), data = data_read)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6288  -0.3990  -0.3240  -0.2678   2.7604
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.213e+00  9.562e-02 -44.062  < 2e-16 ***
## age                2.103e-02  3.517e-03   5.979 2.24e-09 ***
## male               4.139e-01  4.175e-02   9.914  < 2e-16 ***
## friend_cnt        -4.584e-04  2.972e-04  -1.543 0.122942
## avg_friend_age     2.369e-02  4.637e-03   5.108 3.25e-07 ***
## avg_friend_male    1.047e-01  6.555e-02   1.597 0.110222
## subscriber_friend  9.719e-01  4.211e-02  23.080  < 2e-16 ***
## friend_country_cnt 1.401e-02  3.646e-03   3.843 0.000122 ***
## songsListened      6.152e-06  5.212e-07  11.805  < 2e-16 ***
## lovedTracks        6.148e-04  4.828e-05  12.734  < 2e-16 ***
## posts              1.074e-04  9.027e-05   1.189 0.234260
## playlists          6.467e-02  1.310e-02   4.938 7.89e-07 ***
## shouts             7.416e-05  6.476e-05   1.145 0.252113
## tenure            -4.929e-03  1.024e-03  -4.812 1.49e-06 ***
## good_country      -3.939e-01  4.077e-02  -9.661  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

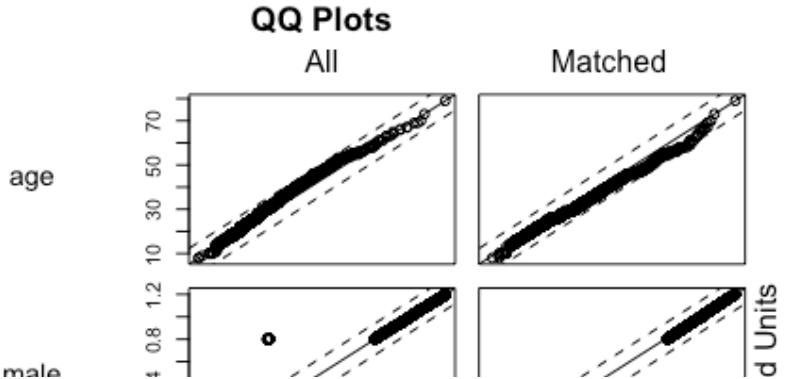
```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24537  on 43826  degrees of freedom
## Residual deviance: 22198  on 43812  degrees of freedom
## AIC: 22228
##
## Number of Fisher Scoring iterations: 5
mylogit_reduced<-glm(adopter~ +age + male + avg_friend_age + subscriber_friend +
friend_country_cnt
                + songsListened + lovedTracks + playlists + tenure +
good_country,data=data_read,family=binomial())

exp(coef(mylogit_reduced))
##        (Intercept)                age               male
##         0.01555748         1.02062333         1.50543746
##     avg_friend_age  subscriber_friend friend_country_cnt
##         1.02533301         2.66293682         1.01067806
##      songsListened        lovedTracks          playlists
##         1.00000631         1.00062174         1.06678846
##            tenure       good_country
##         0.99521304         0.67282465
prs_df <- data.frame(pr_score = predict(mylogit_reduced, type = "response"),
                adopter=mylogit_reduced$model$adopter)
head(prs_df)
##     pr_score adopter
## 1 0.02733510       0
## 2 0.05240358       0
## 3 0.06005164       0
## 4 0.09991419       0
## 5 0.05942411       0
## 6 0.07701414       0
head(mylogit_reduced$model)
##   adopter age male avg_friend_age subscriber_friend friend_country_cnt
## 1       0  22    0       22.57143                 0                  1
## 2       0  35    0       28.00000                 0                  2
## 3       0  27    1       23.00000                 0                  1
## 4       0  21    0       22.94737                 1                  7
## 5       0  24    0       22.28302                 0                  9
## 6       0  21    1       25.00000                 0                  1
##   songsListened lovedTracks playlists tenure good_country
## 1          9687         194         1     59            1
## 2             0           0         0     35            0
## 3           508           0         1     42            0
## 4          1357          32         0     25            0
## 5         89984          20         0     67            0
## 6        124547          10         1     53            1
# Users whoes friends are subscribers are more likely to become premium users.
# According to odd ratios, one more friend who are paid users the user has,
# odds ratio goes up to 2.66 times

# The model calculate the propensity score for each student.
# It is the student's predicted probability of being treated.

#plot histograms of the estimated propensity scores by treatment status

labs <- paste("Actual subscriber_friend_cnt", c("SFC > 0", "SFC = 0"))

prs_df %>%
 mutate(adopter = ifelse(adopter == 1, labs[1], labs[2])) %>%
 ggplot(aes(x = pr_score)) +
 geom_histogram(color = "white") +
 facet_wrap(~adopter) +
 xlab("Probability of being an adopter") +
 theme_bw()
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

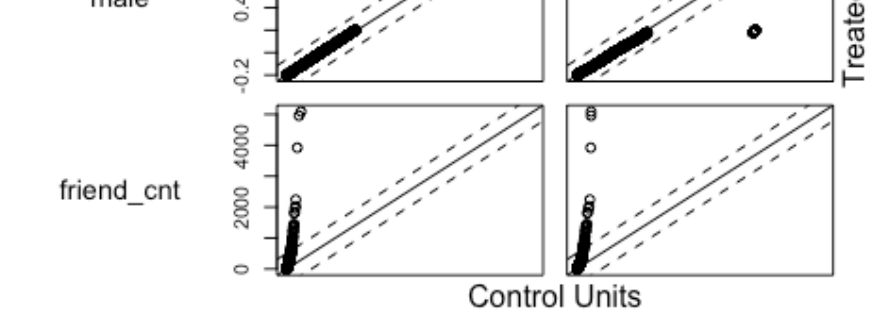


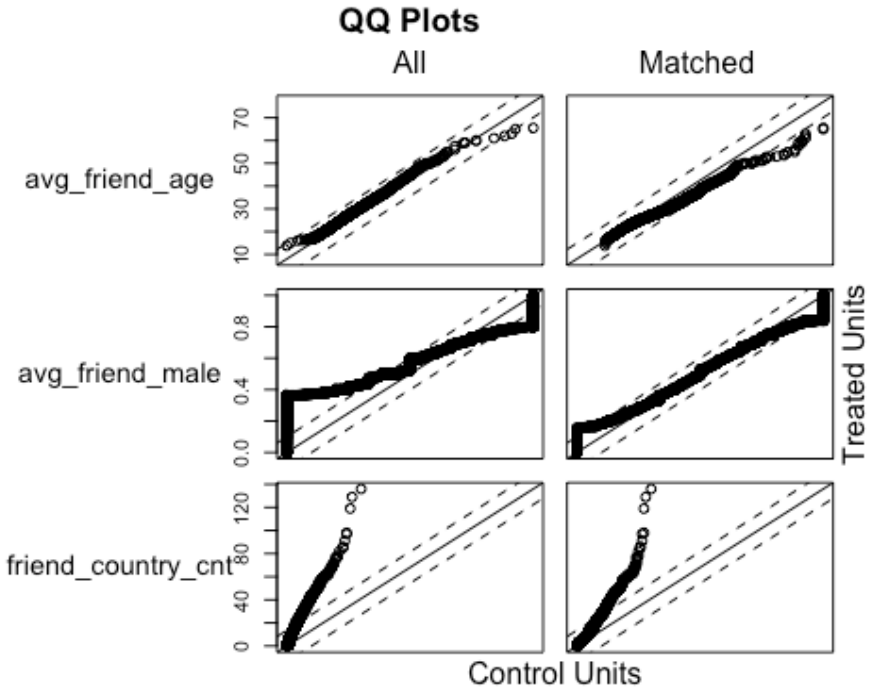


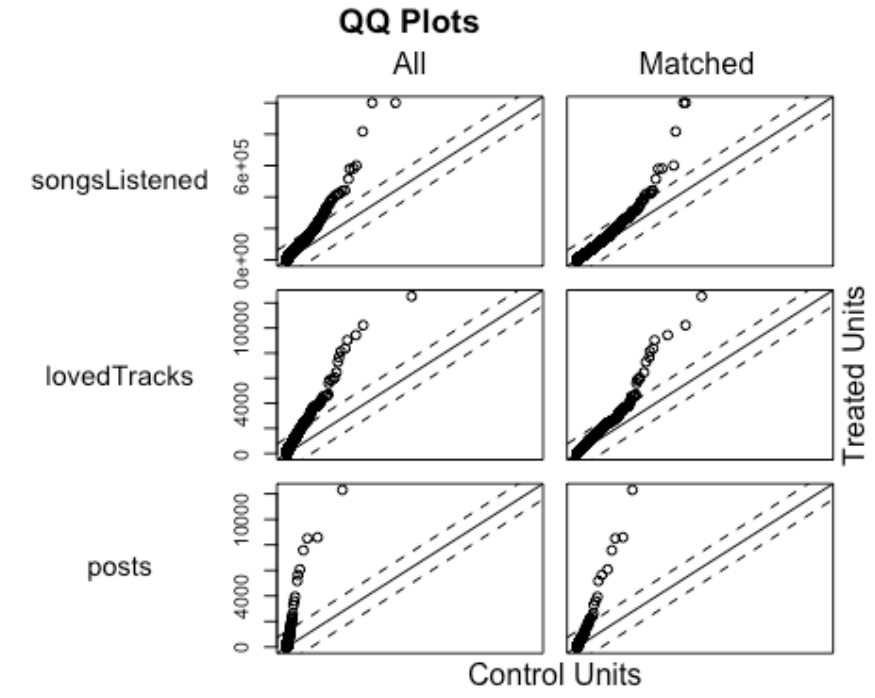

```
#PSM
# The method we use below is to find pairs of observations that have very similar propensity scores,
# but that differ in their treatment status. We use the package MatchIt for this.
# This package estimates the propensity score in the background and then matches observations
based
# on the method of choice ("nearest" in this case).


adopter_nomiss <- data_read %>%  # MatchIt does not allow missing values
 select(adopter, one_of(adopter_cov)) %>%
 na.omit()


mod_match_adopt <- matchit(subscriber_friend ~ age + male + friend_cnt + avg_friend_age +
avg_friend_male
                    + friend_country_cnt+
songsListened+lovedTracks+posts+playlists+shouts+tenure+good_country
                    , method = "nearest", data =data_read)
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
# How successful the matching was using summary(mod_match) and plot(mod_match)

summary(mod_match_adopt)
##
## Call:
## matchit(formula = subscriber_friend ~ age + male + friend_cnt +
##     avg_friend_age + avg_friend_male + friend_country_cnt + songsListened +
##     lovedTracks + posts + playlists + shouts + tenure + good_country,
##     data = data_read, method = "nearest")
##
## Summary of balance for all data:
##                    Means Treated Means Control SD Control  Mean Diff
## distance                  0.4635       0.1550     0.1436     0.3086
## age                      25.3732      23.7476     6.2245     1.6256
## male                      0.6363       0.6288     0.4831     0.0074
## friend_cnt               54.0210      10.4313    15.2769    43.5896
## avg_friend_age           25.3904      23.7614     5.0577     1.6291
## avg_friend_male           0.6358       0.6131     0.3343     0.0227
## friend_country_cnt        9.3856       2.7251     3.1024     6.6606
## songsListened         33735.6404   14602.2205 23214.2898 19133.4199
## lovedTracks             225.3647      65.2137   181.4812   160.1510
## posts                    20.5230       2.5434    33.7947    17.9796
## playlists                 0.7441       0.5295     0.9673     0.2146
## shouts                  101.8195      16.4230    79.7381    85.3965
## tenure                   46.5487      43.2027    19.7212     3.3460
## good_country              0.3433       0.3547     0.4784    -0.0114
##                    eQQ Med  eQQ Mean     eQQ Max
## distance            0.2506     0.3086      0.6840
## age                 1.0000     1.6296      5.0000
## male                0.0000     0.0074      1.0000
## friend_cnt         22.0000    43.5838   4794.0000
## avg_friend_age      1.5909     1.6369     11.5000
## avg_friend_male     0.0738     0.0958      0.3636
## friend_country_cnt  5.0000     6.6598     95.0000
## songsListened   15471.0000 19126.1623 653702.0000
## lovedTracks        65.0000   159.9562   6343.0000
## posts               0.0000    17.8829   9535.0000
## playlists           0.0000     0.2092     26.0000
```

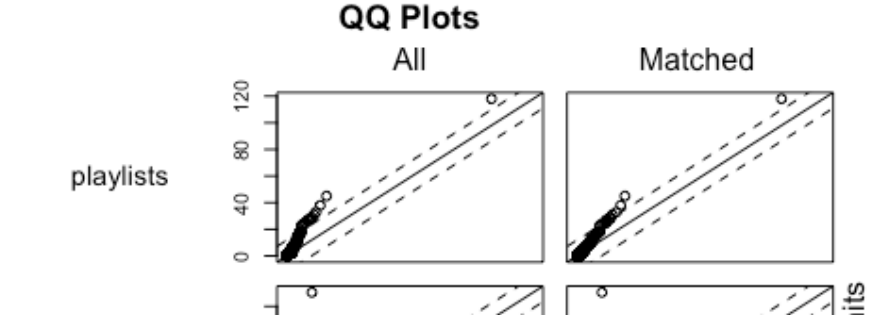
```
## shouts               15.0000    85.1764  59168.0000
## tenure                3.0000     3.3473     10.0000
## good_country          0.0000     0.0114      1.0000
##
##
## Summary of balance for matched data:
##                    Means Treated Means Control SD Control Mean Diff
## distance                  0.4635        0.3040     0.1913    0.1596
## age                      25.3732       26.3324     7.9056   -0.9592
## male                      0.6363        0.6576     0.4745   -0.0214
## friend_cnt               54.0210       21.4666    23.5251   32.5544
## avg_friend_age           25.3904       26.5572     6.7320   -1.1668
## avg_friend_male           0.6358        0.6551     0.2643   -0.0193
## friend_country_cnt        9.3856        5.0914     4.6473    4.2942
## songsListened         33735.6404    27360.8630 33892.7804 6374.7775
## lovedTracks             225.3647      134.5440   299.1995   90.8206
## posts                    20.5230        6.2773    60.2598   14.2456
## playlists                 0.7441        0.6723     1.4015    0.0718
## shouts                  101.8195       37.2362   138.8781   64.5833
## tenure                   46.5487       47.7039    19.0357   -1.1551
## good_country              0.3433        0.3581     0.4795   -0.0149
##                     eQQ Med  eQQ Mean     eQQ Max
## distance             0.1077    0.1596      0.4517
## age                  1.0000    0.9592      7.0000
## male                 0.0000    0.0214      1.0000
## friend_cnt          12.0000   32.5544   4794.0000
## avg_friend_age       0.4376    1.2763     14.0000
## avg_friend_male      0.0158    0.0326      0.1602
## friend_country_cnt   2.0000    4.2942     95.0000
## songsListened     4680.0000 6374.7775 566867.0000
## lovedTracks         38.0000   90.8206   6180.0000
## posts                0.0000   14.2456   9535.0000
## playlists            0.0000    0.1035     22.0000
## shouts              10.0000   64.5833  59168.0000
## tenure               1.0000    1.2995      4.0000
## good_country         0.0000    0.0149      1.0000
##
## Percent Balance Improvement:
##                    Mean Diff. eQQ Med  eQQ Mean  eQQ Max
## distance              48.2930 57.0083   48.2908  33.9658
## age                   40.9972  0.0000   41.1419 -40.0000
## male                -187.9614  0.0000 -187.6712   0.0000
## friend_cnt            25.3162 45.4545   25.3062   0.0000
## avg_friend_age        28.3760 72.4916   22.0309 -21.7391
## avg_friend_male       14.7957 78.6165   65.9532  55.9466
## friend_country_cnt    35.5279 60.0000   35.5203   0.0000
## songsListened         66.6825 69.7499   66.6699  13.2836
## lovedTracks           43.2906 41.5385   43.2216   2.5698
## posts                 20.7676  0.0000   20.3394   0.0000
## playlists             66.5567  0.0000   50.5109  15.3846
## shouts                24.3724 33.3333   24.1770   0.0000
## tenure                65.4771 66.6667   61.1782  60.0000
## good_country         -30.1771  0.0000  -30.3571   0.0000
##
## Sample sizes:
##           Control Treated
## All         34004    9823
## Matched      9823    9823
## Unmatched   24181       0
## Discarded       0       0
plot(mod_match_adopt)
```

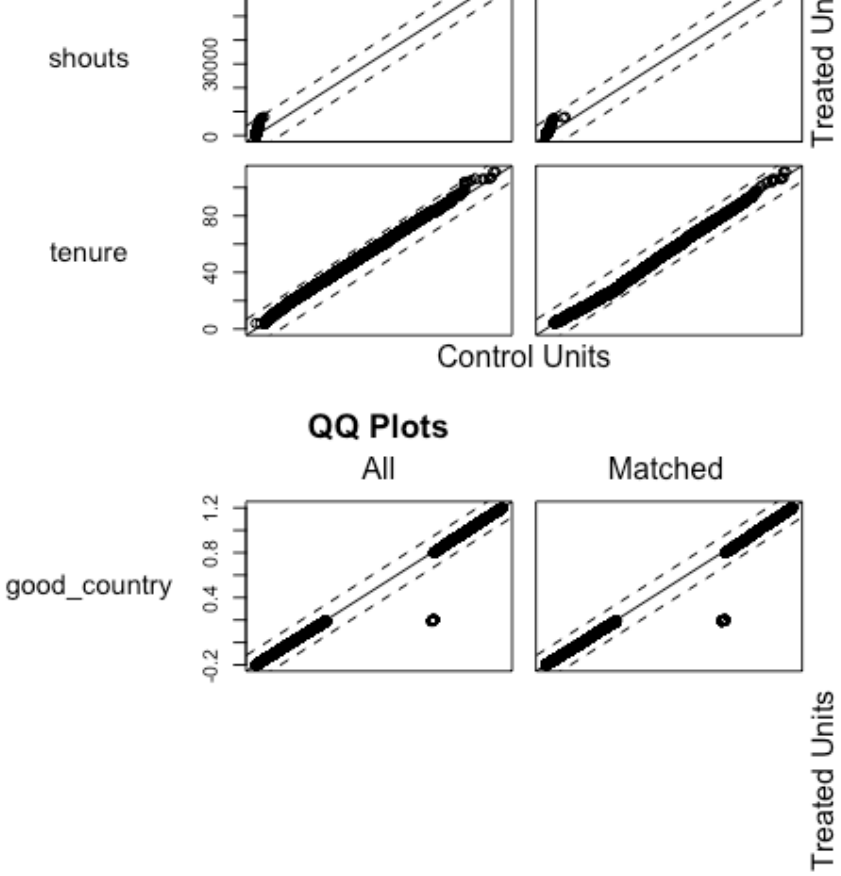



male
friend_cnt
Treated Units
Control Units


QQ Plots
All
Matched
avg_friend_age
avg_friend_male
friend_country_cnt
Treated Units
Control Units

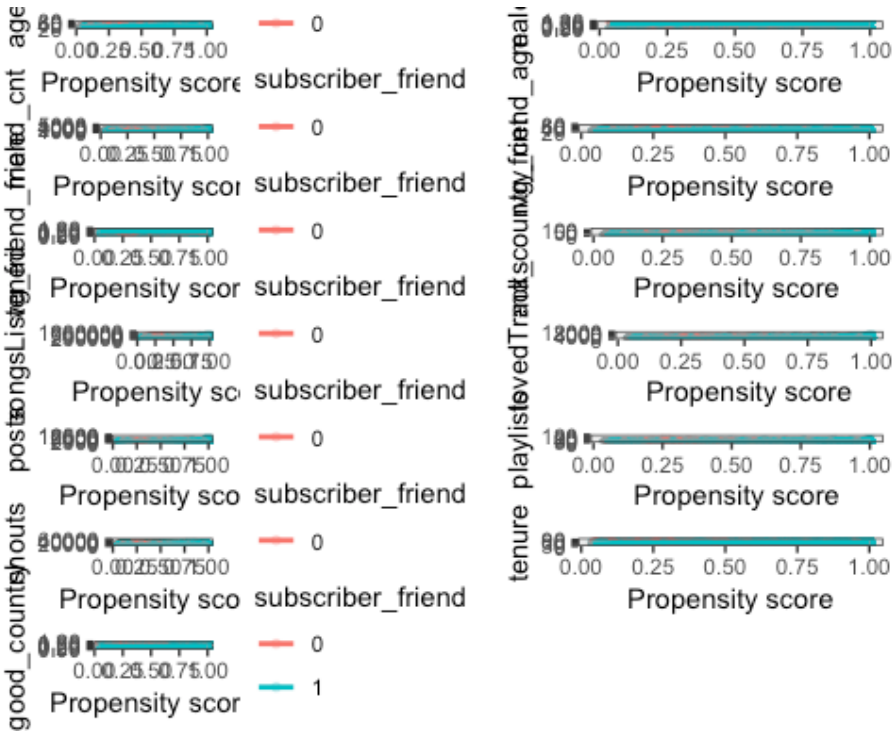

QQ Plots
All
Matched
songsListened
lovedTracks
posts
Treated Units
Control Units


QQ Plots
All
Matched
playlists




```
# To create a dataframe containing only the matched observations, use the match.data() function

dta_m_adopt <- match.data(mod_match_adopt)

dim(dta_m_adopt)
## [1] 19646    19
#Visual Inspection

fn_bal <- function(dta_m_adopt, variable) {
 dta_m_adopt$variable <- dta_m_adopt[, variable]
 dta_m_adopt$subscriber_friend <- as.factor(dta_m_adopt$subscriber_friend)
 support <- c(min(dta_m_adopt$variable), max(dta_m_adopt$variable))
 ggplot(dta_m_adopt, aes(x = distance, y = variable, color = subscriber_friend)) +
  geom_point(alpha = 0.2, size = 1.3) +
  geom_smooth(method = "loess", se = F) +
  xlab("Propensity score") +
  ylab(variable) +
  theme_bw() +
  ylim(support)
}

library(gridExtra)
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
grid.arrange(
 fn_bal(dta_m_adopt, "age"),
 fn_bal(dta_m_adopt, "male") + theme(legend.position = "none"),
 fn_bal(dta_m_adopt, "friend_cnt"),
 fn_bal(dta_m_adopt, "avg_friend_age") + theme(legend.position = "none"),
 fn_bal(dta_m_adopt, "avg_friend_male"),
 fn_bal(dta_m_adopt, "friend_country_cnt") + theme(legend.position = "none"),
```

```
 fn_bal(dta_m_adopt, "songsListened"),
 fn_bal(dta_m_adopt, "lovedTracks") + theme(legend.position = "none"),
 fn_bal(dta_m_adopt, "posts"),
 fn_bal(dta_m_adopt, "playlists") + theme(legend.position = "none"),
 fn_bal(dta_m_adopt, "shouts"),
 fn_bal(dta_m_adopt, "tenure") + theme(legend.position = "none"),
 fn_bal(dta_m_adopt, "good_country"),
 nrow = 7, widths = c(1, 0.8)
)
## Warning: Removed 4 rows containing missing values (geom_smooth).
```



#Difference in means

```
with(dta_m_adopt, t.test(adopter~subscriber_friend))
##
##  Welch Two Sample t-test
##
## data:  adopter by subscriber_friend
## t = -18.938, df = 18060, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.10009352 -0.08131745
## sample estimates:
## mean in group 0 mean in group 1
##      0.08683702      0.17754250
lm_treat1_adopter <- glm(adopter~subscriber_friend, data = dta_m_adopt,family=binomial())
summary(lm_treat1_adopter)
##
## Call:
## glm(formula = adopter ~ subscriber_friend, family = binomial(),
##     data = dta_m_adopt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6252  -0.6252  -0.4262  -0.4262   2.2108
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.35288    0.03583  -65.67   <2e-16 ***
## subscriber_friend  0.81979    0.04451   18.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15345  on 19645  degrees of freedom
## Residual deviance: 14986  on 19644  degrees of freedom
## AIC: 14990
##
```

```
## Number of Fisher Scoring iterations: 5
exp(coef(lm_treat1_adopter))
##       (Intercept) subscriber_friend
##        0.09509476        2.27003359
lm_treat2_adopter <- glm(adopter~age + male + friend_cnt + avg_friend_age + avg_friend_male +
friend_country_cnt+ subscriber_friend+
songsListened+lovedTracks+posts+playlists+shouts+tenure+good_country, data =
dta_m_adopt,family = binomial())
summary(lm_treat2_adopter)
##
## Call:
## glm(formula = adopter ~ age + male + friend_cnt + avg_friend_age +
##     avg_friend_male + friend_country_cnt + subscriber_friend +
##     songsListened + lovedTracks + posts + playlists + shouts +
##     tenure + good_country, family = binomial(), data = dta_m_adopt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2240  -0.5668  -0.4562  -0.3697   2.5257
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.371e+00  1.261e-01 -26.727  < 2e-16 ***
## age                1.419e-02  4.069e-03   3.488 0.000486 ***
## male               3.040e-01  4.899e-02   6.205 5.49e-10 ***
## friend_cnt        -2.002e-04  2.793e-04  -0.717 0.473545
## avg_friend_age     1.304e-02  5.349e-03   2.438 0.014757 *
## avg_friend_male    6.007e-02  9.252e-02   0.649 0.516196
## friend_country_cnt  7.277e-03  3.649e-03   1.995 0.046088 *
## subscriber_friend   7.293e-01  4.682e-02  15.578  < 2e-16 ***
## songsListened      4.255e-06  5.301e-07   8.027 9.97e-16 ***
## lovedTracks        5.211e-04  4.692e-05  11.105  < 2e-16 ***
## posts              1.186e-04  8.891e-05   1.334 0.182212
## playlists          4.465e-02  1.194e-02   3.738 0.000185 ***
## shouts             1.119e-04  7.454e-05   1.502 0.133156
## tenure            -2.434e-03  1.217e-03  -1.999 0.045556 *
## good_country      -3.695e-01  4.802e-02  -7.695 1.42e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15345  on 19645  degrees of freedom
## Residual deviance: 14493  on 19631  degrees of freedom
## AIC: 14523
##
## Number of Fisher Scoring iterations: 5
exp(coef(lm_treat2_adopter))
##       (Intercept)               age              male
##        0.03435635        1.01429580        1.35522294
##        friend_cnt    avg_friend_age   avg_friend_male
##        0.99979986        1.01312796        1.06190965
## friend_country_cnt  subscriber_friend      songsListened
##        1.00730402        2.07357751        1.00000426
##       lovedTracks             posts         playlists
##        1.00052122        1.00011861        1.04566035
##            shouts            tenure      good_country
##        1.00011195        0.99756942        0.69108548
```