Social Inclusion Analysis

Yining Hua

01/27/2021

```
dat <- read_csv("preliminary_data.csv")

col_names <- c("job","marriage", "group","gender","has_local_child",
              "from_rural")
dat[,col_names] <- lapply(dat[,col_names] , factor)

dat$money.left <- dat$income - dat$expence

## Normalize and standardize money left
dat$money.left <- (dat$money.left - mean(dat$money.left, na.rm = TRUE)) /
                    sd(dat$money.left, na.rm = TRUE)
dat$participant <- as.character(dat$participant)

dat$migration.scale <- as.factor(dat$migration.scale)
dat$education.group <- as.factor(dat$education)
```

```
head(dat)
```

```
## # A tibble: 6 x 29
##   participant gender education marriage migration.scale age_group expence income
##   <chr>       <fct>  <chr>     <fct>    <fct>           <chr>       <dbl>  <dbl>
## 1 0           female highscho~ 1        interstate      23-64       10000     NA
## 2 2           female highscho~ 1        interstate      23-64       40000     NA
## 3 4           female junior c~ 1        interstate      23-64        9000     NA
## 4 6           female highscho~ 1        interstate      23-64       18000 -90000
## 5 12          female highscho~ 1        intercity       23-64        6500 -16000
## 6 13          male   highscho~ 1        intercity       23-64        4000 -13000
## # ... with 21 more variables: worked_before5.1 <dbl>, job <fct>,
## #   has_local_child <fct>, housing_type <chr>, from_rural <fct>,
## #   time_stayed <dbl>, hangouts <chr>, willing.to.movein <chr>, pos_stay <dbl>,
## #   neg_stay <dbl>, diabete.or.hypertension <chr>, group <fct>,
## #   participated.in.group.activity <dbl>, like.current.city <dbl>,
## #   natives.like.me <dbl>, natives.lookdown.me <dbl>,
## #   previous.customs.better <dbl>, i.am.native <dbl>, insuranced <dbl>, ...
```

```
# dat$natives_inclusion <- dat$natives.like.me-dat$natives.lookdown.me
dat$city_inclusion <- dat$like.current.city - dat$previous.customs.better +
                      dat$i.am.native + dat$natives.like.me -
                      dat$natives.lookdown.me
# dat$tendency.livehere <- dat$willing.to.movein + dat$willing.to.stay
```

```
dat
```

```
## # A tibble: 65,432 x 30
##    participant gender education      marriage migration.scale age_group expence
##    <chr>       <fct>  <chr>          <fct>    <fct>           <chr>       <dbl>
##  1 0           female highschool     1        interstate      23-64       10000
##  2 2           female highschool     1        interstate      23-64       40000
##  3 4           female junior college 1        interstate      23-64        9000
##  4 6           female highschool     1        interstate      23-64       18000
##  5 12          female highschool     1        intercity       23-64        6500
##  6 13          male   highschool     1        intercity       23-64        4000
##  7 16          male   highschool     1        interstate      23-64        2000
##  8 18          female midschool      1        intercounty     23-64        5000
##  9 23          female midschool      0        intercity       23-64        4200
## 10 25          male   highschool     1        intercity       23-64        6000
## # ... with 65,422 more rows, and 23 more variables: income <dbl>,
## #   worked_before5.1 <dbl>, job <fct>, has_local_child <fct>,
## #   housing_type <chr>, from_rural <fct>, time_stayed <dbl>, hangouts <chr>,
## #   willing.to.movein <chr>, pos_stay <dbl>, neg_stay <dbl>,
## #   diabete.or.hypertension <chr>, group <fct>,
## #   participated.in.group.activity <dbl>, like.current.city <dbl>,
## #   natives.like.me <dbl>, natives.lookdown.me <dbl>, ...
```

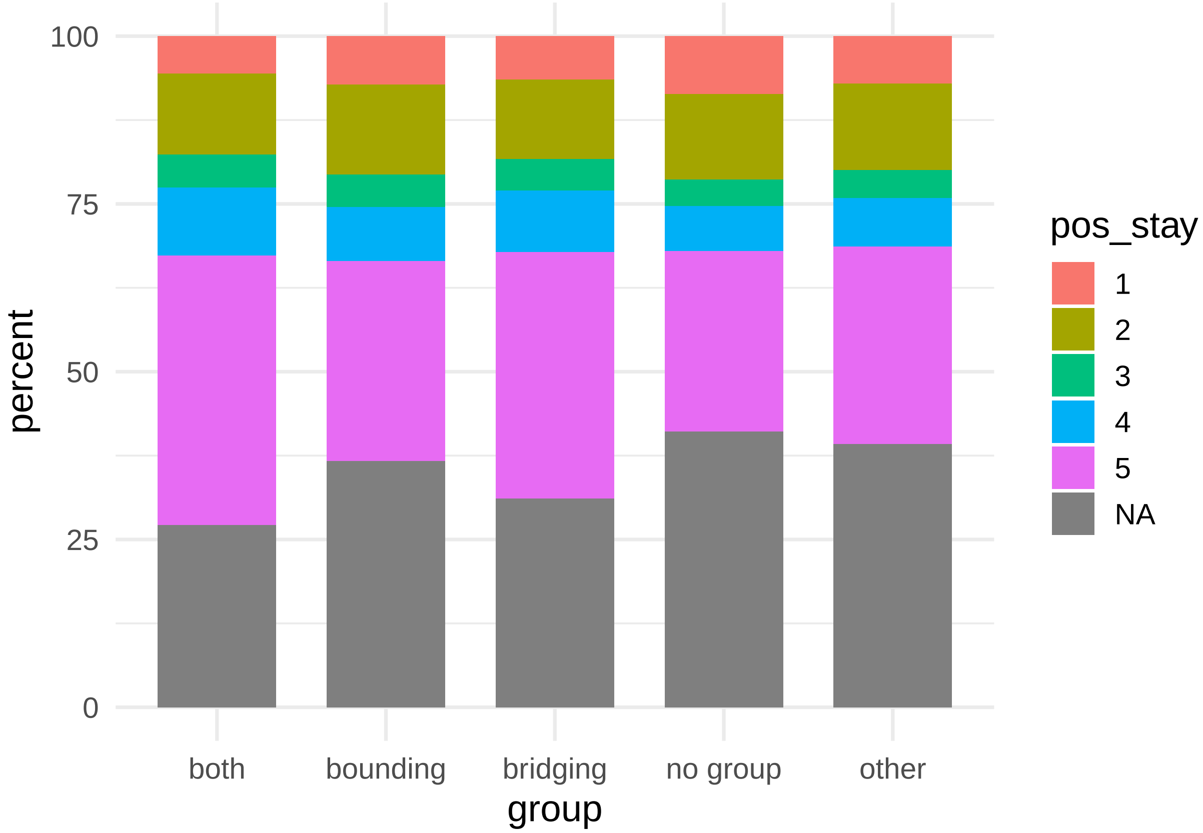
```
## regroup education
dat$education.group <- NA
dat$education.group[dat$education == "no education"] <- "low"
dat$education.group[dat$education == "primary school"] <- "low"
dat$education.group[dat$education == "midschool"] <- "middle"
dat$education.group[dat$education == "highschool"] <- "middle"
dat$education.group[dat$education =="junior college"] <- "middle"
dat$education.group[dat$education == "college"] <- "high"
dat$education.group[dat$education == "grad"] <- "high"
```

```
## regroup ethnicity
#dat$ethnicity.group <- "other"
#dat$ethnicity.group[dat$ethnicity == 1] <- "han"
```

```
cbPalette <- c("#e61212", "#ffb300", "#22ff00", "#0015ff", "#00fbff")
d2 <- dat %>%
  group_by(group, pos_stay) %>%
  summarise(count = n()) %>%
  mutate(perc = count/sum(count))
```

```
## ‘summarise()‘ has grouped output by ’group’. You can override using the ‘.groups‘ argument.
```

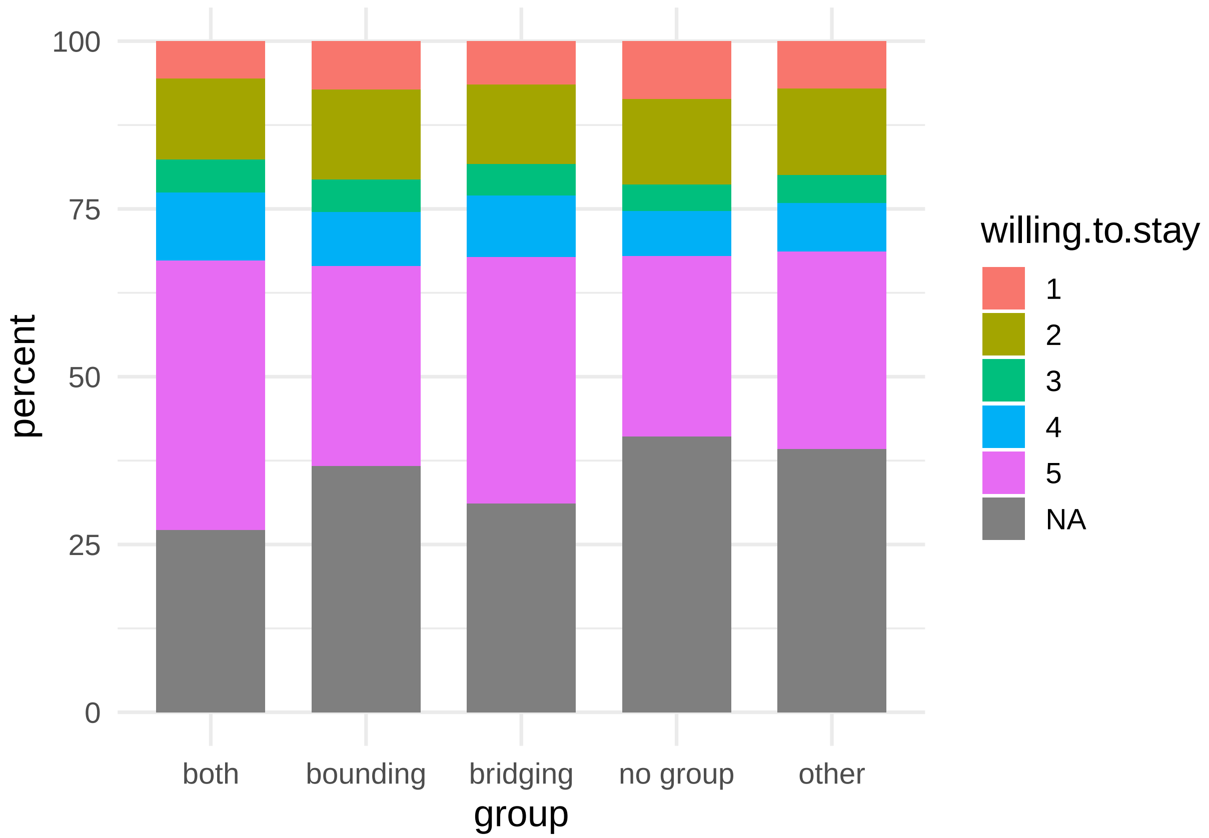
```
ggplot(d2, aes(x = factor(group), y = perc*100, fill = factor(pos_stay))) +
  geom_bar(stat="identity", width = 0.7) +
  labs(x = "group", y = "percent", fill = "pos_stay") +
  theme_minimal(base_size = 14)
```



```
cbPalette <- c("#e61212", "#ffb300", "#22ff00", "#0015ff", "#00fbff")
d2 <- dat %>%
  group_by(group, pos_stay) %>%
  summarise(count = n()) %>%
  mutate(perc = count/sum(count))
```

```
## ‘summarise()‘ has grouped output by ’group’. You can override using the ‘.groups‘ argument.
```

```
ggplot(d2, aes(x = factor(group), y = perc*100, fill = factor(pos_stay))) +
  geom_bar(stat="identity", width = 0.7) +
  labs(x = "group", y = "percent", fill = "willing.to.stay") +
  theme_minimal(base_size = 14)
```



```
dat$group <- relevel(dat$group, ref = "no group")
dat$job <- relevel(dat$job, ref = "unstable job")
dat$education.group <- relevel(as.factor(dat$education.group), ref = "low")
dat$migration.scale <- relevel(as.factor(dat$migration.scale),
                                              ref = "intercounty")
dat$age_group <- relevel(as.factor(dat$age_group), ref = "15-22")
```

```
library(broom)

mod1 <- glm(pos_stay ~ group, data=dat)
mod2 <- glm(neg_stay ~ group, data=dat)
```

```
summary(mod1)
```

```
##
## Call:
## glm(formula = pos_stay ~ group, data = dat)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.922  -1.518   1.078   1.162   1.482
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.51761    0.01258 279.558   < 2e-16 ***
## groupboth      0.40422    0.01859  21.739  < 2e-16 ***
## groupbounding  0.10977    0.02275   4.825 1.41e-06 ***
## groupbridging  0.32039    0.01956  16.377  < 2e-16 ***
## groupother     0.12487    0.05557   2.247   0.0246 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 2.220365)
##
##     Null deviance: 96055  on 42701  degrees of freedom
## Residual deviance: 94803  on 42697  degrees of freedom
##   (22730 observations deleted due to missingness)
## AIC: 155252
##
## Number of Fisher Scoring iterations: 2
```

```
summary(mod2)
```

```
##
## Call:
## glm(formula = neg_stay ~ group, data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3365  -0.3365   0.6635   0.6635   1.0275
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.3365     0.0588  73.754  < 2e-16 ***
## groupboth      -0.3640     0.1164  -3.128  0.00184 **
## groupbounding  -0.2740     0.1221  -2.244  0.02517 *
## groupbridging  -0.2188     0.1074  -2.037  0.04204 *
## groupother     -0.1365     0.4726  -0.289  0.77283
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 1.099328)
##
##     Null deviance: 738.71  on 663  degrees of freedom
## Residual deviance: 724.46  on 659  degrees of freedom
##   (64768 observations deleted due to missingness)
## AIC: 1954.2
##
## Number of Fisher Scoring iterations: 2
```

```
mod3 <- glm(pos_stay ~ group + city_inclusion + money.left +
            education.group + age_group + marriage + job + has_local_child +
            gender + migration.scale + housing_type + from_rural + time_stayed,
              data=dat)
summary(mod3)
```

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

```
## glm(formula = pos_stay ~ group + city_inclusion + money.left +
##     education.group + age_group + marriage + job + has_local_child +
##     gender + migration.scale + housing_type + from_rural + time_stayed,
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0590  -1.1222  -0.0012   1.1271   3.7872
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.622385   0.097119  16.705  < 2e-16 ***
## groupboth                     0.182013   0.036943   4.927 8.50e-07 ***
## groupbounding                 0.080550   0.043054   1.871 0.061387 .
## groupbridging                 0.176446   0.037965   4.648 3.40e-06 ***
## groupother                    0.042032   0.107918   0.389 0.696929
## city_inclusion                0.112695   0.004952  22.755  < 2e-16 ***
## money.left                    0.046394   0.017164   2.703 0.006884 **
## education.grouphigh           0.760090   0.066945  11.354  < 2e-16 ***
## education.groupmiddle         0.320021   0.049709   6.438 1.27e-10 ***
## age_group>=65                 0.514759   0.222777   2.311 0.020873 *
## age_group23-64                0.408641   0.049401   8.272  < 2e-16 ***
## marriage1                     0.167028   0.035499   4.705 2.57e-06 ***
## jobstable job                 0.020194   0.057001   0.354 0.723148
## has_local_child1              0.507590   0.044022  11.530  < 2e-16 ***
## gendermale                   -0.118090   0.028754  -4.107 4.04e-05 ***
## migration.scaleintercity      0.036382   0.041869   0.869 0.384901
## migration.scaleinterstate -0.141910   0.040240  -3.527 0.000423 ***
## housing_typeownership         0.837189   0.049736  16.833  < 2e-16 ***
## from_rural1                  -0.163319   0.031939  -5.114 3.22e-07 ***
## time_stayed                   0.048699   0.002221  21.930  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 1.849145)
##
##     Null deviance: 23794  on 9681  degrees of freedom
## Residual deviance: 17866  on 9662  degrees of freedom
##   (55750 observations deleted due to missingness)
## AIC: 33450
##
## Number of Fisher Scoring iterations: 2
```

We took out the has_local_child co-variate in this model because no one had the variable to be 1 in this group.

```
mod4 <- glm(neg_stay ~ group + city_inclusion + money.left +
            education.group + age_group + marriage + job +
            gender + migration.scale + housing_type + from_rural + time_stayed,
              data=dat)
summary(mod4)
```

```
##
## Call:
## glm(formula = neg_stay ~ group + city_inclusion + money.left +
##     education.group + age_group + marriage + job + gender + migration.scale +
##     housing_type + from_rural + time_stayed, data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3327  -0.6353   0.2228   0.7006   2.0538
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                5.42330    0.50593  10.719  < 2e-16 ***
## groupboth                 -0.35654    0.25207  -1.414    0.159
## groupbounding              0.01181    0.23830   0.050    0.961
## groupbridging             -0.06422    0.21556  -0.298    0.766
## groupother                -0.94281    0.65012  -1.450    0.149
## city_inclusion             0.03870    0.02395   1.615    0.108
## money.left                -0.20122    0.13813  -1.457    0.147
## education.grouphigh        0.24202    0.36629   0.661    0.510
## education.groupmiddle     -0.12655    0.20336  -0.622    0.535
## age_group>=65              0.03582    0.84418   0.042    0.966
## age_group23-64            -0.32355    0.29227  -1.107    0.270
## marriage1                 -0.02004    0.18950  -0.106    0.916
## jobstable job             -0.30921    0.28347  -1.091    0.277
## gendermale                -0.17401    0.16986  -1.024    0.307
## migration.scaleintercity  -0.10900    0.25552  -0.427    0.670
## migration.scaleinterstate -0.22628    0.23187  -0.976    0.330
## housing_typeownership     -0.16233    0.44256  -0.367    0.714
## from_rural1               -0.16656    0.19497  -0.854    0.394
## time_stayed               -0.04987    0.01119  -4.458 1.46e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 1.146804)
##
##     Null deviance: 260.35  on 195  degrees of freedom
## Residual deviance: 202.98  on 177  degrees of freedom
##   (65236 observations deleted due to missingness)
## AIC: 603.09
##
## Number of Fisher Scoring iterations: 2
```