EGC Data Test

Preksha Jain

3/8/2020

Data Preparation

```
library("rio") # loading library to convert .dta file to .csv
```

```
## Warning: package 'rio' was built under R version 3.6.3
```

```
getwd()
```

```
## [1] "C:/Users/preks/Documents/Yale/STATATests/EGC_Stata_Test_2020"
```

```
setwd("C:/Users/preks/Documents/Yale/STATATests/EGC_Stata_Test_2020")
convert("endline.dta", "endline.csv")

# Reading endline data into R
endline <- read.csv("endline.csv", header = T)

# Browsing and understanding variables and data structure
str(endline)
```

```
## 'data.frame':    4160 obs. of  7 variables:
##  $ hhid                : int  86 147 179 192 261 268 294 353 450 500 ...
##  $ group_id            : int  3 96 4 76 14 96 14 13 76 134 ...
##  $ totformalborrow_24  : Factor w/ 414 levels ".","10000","100000",..: 33
414 305 60 414 33 414 130 72 169 ...
##  $ totinformalborrow_24: Factor w/ 347 levels ".","10000","100000",..: 294
180 342 347 347 347 347 115 2 347 ...
##  $ hhinc               : Factor w/ 803 levels "100","1000","10000",..: 802
36 511 802 622 341 308 329 161 760 ...
##  $ hhnomembers         : int  4 4 5 2 7 5 2 5 6 6 ...
##  $ survey_round        : Factor w/ 3 levels "Endline I","Endline II",..: 2
2 2 2 1 2 1 2 2 2 ...
```

```
dim(endline)
```

```
## [1] 4160    7
```

```
summary(endline)
```

```
##       hhid           group_id     totformalborrow_24 totinformalborrow_24
##  Min.   :    86   Min.   :  1.0   None   :1221       None   :1539
##  1st Qu.: 78535   1st Qu.: 76.0   20000  : 242       10000  : 212
##  Median :114348   Median :133.0   50000  : 133       5000   : 134
##  Mean   :105520   Mean   :113.2   30000  : 128       50000  : 117
##  3rd Qu.:125841   3rd Qu.:159.0   25000  : 120       20000  : 115
```

```
##  Max.   :185878   Max.   :183.0   15000  : 113       30000  : 103
##                                   (Other):2203       (Other):1940
##      hhinc       hhnomembers          survey_round
##  None   : 244   Min.   : 1.000   Endline I  : 641
##  10000  : 162   1st Qu.: 3.000   Endline II :2873
##  3000   : 156   Median : 4.000   Endline III: 646
##  1000   : 135   Mean   : 4.514
##  5000   : 133   3rd Qu.: 6.000
##  6000   : 130   Max.   :16.000
##  (Other):3200
```

```
# Replacing "None" with "0" in new debt and income variables, simultaneously
changing variables' class to numeric
endline$new_totformbor_24 <- as.numeric(gsub("None", "0", endline$totformalbo
rrow_24))
```

```
## Warning: NAs introduced by coercion
```

```
endline$newtotinformbor_24 <- as.numeric(gsub("None", "0", endline$totinforma
lborrow_24))
```

```
## Warning: NAs introduced by coercion
```

```
endline$new_hhinc <- as.numeric(gsub("None", "0", endline$hhinc))
```

```
## Warning: NAs introduced by coercion
```

```
# Checking dimensions and summary stats
dim(endline)
```

```
## [1] 4160   10
```

```
summary(endline)
```

```
##       hhid           group_id     totformalborrow_24 totinformalborrow_24
##  Min.   :    86   Min.   :  1.0   None   :1221       None   :1539
##  1st Qu.: 78535   1st Qu.: 76.0   20000  : 242       10000  : 212
##  Median :114348   Median :133.0   50000  : 133       5000   : 134
##  Mean   :105520   Mean   :113.2   30000  : 128       50000  : 117
##  3rd Qu.:125841   3rd Qu.:159.0   25000  : 120       20000  : 115
##  Max.   :185878   Max.   :183.0   15000  : 113       30000  : 103
##                                   (Other):2203       (Other):1940
##      hhinc       hhnomembers          survey_round  new_totformbor_24
##  None   : 244   Min.   : 1.000   Endline I  : 641   Min.   :      0
##  10000  : 162   1st Qu.: 3.000   Endline II :2873   1st Qu.:      0
##  3000   : 156   Median : 4.000   Endline III: 646   Median :  30000
##  1000   : 135   Mean   : 4.514                      Mean   :  64382
##  5000   : 133   3rd Qu.: 6.000                      3rd Qu.:  75000
##  6000   : 130   Max.   :16.000                      Max.   :3690000
##  (Other):3200                                       NA's   :4
##  newtotinformbor_24   new_hhinc
##  Min.   :      0    Min.   :      0
```

```
##  1st Qu.:      0    1st Qu.:   2850
##  Median :  10000    Median :   6000
##  Mean   :  40921    Mean   :  11809
##  3rd Qu.:  45000    3rd Qu.:  11000
##  Max.   :1120000    Max.   :4000000
##  NA's   :4          NA's   :4
```

```
# Inferences about the financial status of households - checking if the mean/
median income for HHs which borrow more formally differs from those those bor
row more informally
summary(endline$new_hhinc[endline$new_totformbor_24 > endline$newtotinformbor
_24])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0    3185    7000   14421   12644 4000000       5
```

```
summary(endline$new_hhinc[endline$new_totformbor_24 < endline$newtotinformbor
_24])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0    3000    6000    9992   10000 1062000       5
```

```
summary(endline$new_totformbor_24[endline$new_totformbor_24 > endline$newtoti
nformbor_24])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2000   30000   60000  107313  125000 3690000       4
```

```
summary(endline$newtotinformbor_24[endline$new_totformbor_24 < endline$newtot
informbor_24])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2000   20000   50000   90327  110000 1120000       4
```

```
summary(endline$newtotinformbor_24[endline$new_totformbor_24 > endline$newtot
informbor_24])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       0    3000   17872   19500  650000       4
```

```
summary(endline$new_totformbor_24[endline$new_totformbor_24 < endline$newtoti
nformbor_24])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       0    3000   23520   30000  405000       4
```

```
# It seems preliminarily that there is indeed a difference between the mean i
ncome of households which borrow more formally vs. informally, i.e. richer ho
useholds can afford/have access to formal sources of lending as compared to p
oorer households which have to rely on more informal sources of lending. It a
lso seems like the ticket size for formal vs. informal borrowing is higher fo
r those who borrow dominantly from the respective sources, however, not so mu
```

```
ch if we look at non-dominant sources. These preliminary inferences can be ma
de more concrete through a t-test.

# Creating a function to replace outliers (values greater than three s.d.'s o
f the mean) with cutoff value -> Top coding
outlierReplace <- function(x){
  cutoff <- mean(x[!is.na(x)]) + 3*sqrt(var(x[!is.na(x)]))
  x[x>cutoff] <- cutoff
  return(x)
}

# Applying the function (top-coding) to debt and income variables
endline$new_hhinc <- outlierReplace(endline$new_hhinc)
endline$new_totformbor_24 <- outlierReplace(endline$new_totformbor_24)
endline$newtotinformbor_24 <- outlierReplace(endline$newtotinformbor_24)

# It is not possible to label variables in R so I have just replaced them
# It is important to top-code income and debt variabes since we don't want ou
tliers in the data to drive and show treatment effects - it is to make the mo
del more robust.
# Other checks could be to test for "good variation" in our data, and retain
only those variables that satisfy a given criterion. Another could be to chec
k for missing values and how to handle them.

# Creating a new variable that captures total borrowed amount in the last 24
months (sum of formal and informal borrowing in the L24M)
endline$new_totbor_24 <- rowSums(cbind(endline$new_totformbor_24, endline$new
totinformbor_24), na.rm = T)

# Loading the treatment status data into R
treatment_status <- read.csv("treatment_status.csv", header = T)

# Analyzing the data structure
str(treatment_status)

## 'data.frame':    101 obs. of  3 variables:
##  $ pair_id : int  34 31 14 31 5 1 15 21 17 2 ...
##  $ group_id: int  35 3 96 4 76 14 13 134 122 57 ...
##  $ treated : int  1 1 0 0 0 0 1 0 0 0 ...

# Changing variable classes from integer to factor as appropriate
treatment_status[,c(1:3)] <- lapply(treatment_status[,c(1:3)], as.factor)

# Re-analyzing data structure after modifying variable classes
str(treatment_status)

## 'data.frame':    101 obs. of  3 variables:
##  $ pair_id : Factor w/ 50 levels "1","2","4","5",..: 29 26 12 26 4 1 13 18
15 2 ...
##  $ group_id: Factor w/ 101 levels "1","2","3","4",..: 14 3 40 4 27 6 5 54
```

```
47 21 ...
##  $ treated : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 2 1 1 1 ...

# Verifying the information given in the question about 51 control groups and
50 treatment groups
table(treatment_status$treated)

##
##  0  1
## 51 50

# Merging the treatment_status data with the endline data by using the common
column "group_id"
endline_merged <- merge(endline, treatment_status, by = "group_id")

# Checking the data structure of the merged dataset
str(endline_merged)

## 'data.frame':    4160 obs. of  13 variables:
##  $ group_id            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hhid                : int  129607 130298 130409 130413 130299 130290 13
0310 130518 130430 130304 ...
##  $ totformalborrow_24  : Factor w/ 414 levels ".","10000","100000",..: 345
414 414 220 130 169 414 46 399 169 ...
##  $ totinformalborrow_24: Factor w/ 347 levels ".","10000","100000",..: 347
347 179 222 347 2 347 296 179 149 ...
##  $ hhinc               : Factor w/ 803 levels "100","1000","10000",..: 619
611 722 765 349 423 676 561 318 25 ...
##  $ hhnomembers         : int  4 4 6 7 4 4 9 4 8 6 ...
##  $ survey_round        : Factor w/ 3 levels "Endline I","Endline II",..: 2
2 2 2 2 2 2 2 2 ...
##  $ new_totformbor_24   : num  66000 0 0 320000 200000 25000 0 130000 90000
25000 ...
##  $ newtotinformbor_24  : num  0 0 30000 40000 0 10000 0 70000 30000 2500 .
..
##  $ new_hhinc           : num  5950 58400 8000 9000 26500 32500 7100 5000 2
400 10500 ...
##  $ new_totbor_24       : num  66000 0 30000 360000 200000 35000 0 200000 1
20000 27500 ...
##  $ pair_id             : Factor w/ 50 levels "1","2","4","5",..: 25 25 25
25 25 25 25 25 25 25 ...
##  $ treated             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 .
..

head(endline_merged)

##   group_id   hhid totformalborrow_24 totinformalborrow_24 hhinc hhnomember
s
## 1        1 129607              66000                 None  5950
4
## 2        1 130298               None                 None 58400
```

```
4
## 3        1 130409               None                30000  8000
6
## 4        1 130413             320000                40000  9000
7
## 5        1 130299             200000                 None 26500
4
## 6        1 130290              25000                10000 32500
4
##   survey_round new_totformbor_24 newtotinformbor_24 new_hhinc new_totbor_2
4
## 1   Endline II             66000                  0      5950         6600
0
## 2   Endline II                 0                  0     58400
0
## 3   Endline II                 0              30000      8000         3000
0
## 4   Endline II            320000              40000      9000        36000
0
## 5   Endline II            200000                  0     26500        20000
0
## 6   Endline II             25000              10000     32500         3500
0
##   pair_id treated
## 1      30       0
## 2      30       0
## 3      30       0
## 4      30       0
## 5      30       0
## 6      30       0
```

```
dim(endline_merged)
```

```
## [1] 4160   13
```

```
# Creating a dummy variable for HHs with per capita daily income below the po
verty line of Rs. 26.995 (2010 PPP conversion of USD 1.90). This dummy takes
the value 1 if below poverty line, 0 if not. As mentioned earlier, it is not
possible to label variables in R.
endline_merged$bpl <- as.factor(ifelse(endline_merged$new_hhinc/endline_merge
d$hhnomembers/30 < 26.995, 1, 0))

# There are 4 missing values reported for HHs which refused to answer the que
stion on household income
endline_merged[is.na(endline_merged$bpl),]
```

```
##      group_id   hhid totformalborrow_24 totinformalborrow_24            hh
inc
## 58          2  21109               None                 None Refuse to ans
wer
## 2315      139 128691             152000                97000 Refuse to ans
```

```
wer
## 2393      142 105091               None                50000 Refuse to ans
wer
## 3161      160 122909               None                 None Refuse to ans
wer
##      hhnomembers survey_round new_totformbor_24 newtotinformbor_24 new_hhi
nc
## 58             2   Endline II                 0                  0
NA
## 2315           8   Endline II            152000             97000
NA
## 2393           3  Endline III                 0             50000
NA
## 3161           6   Endline II                 0                  0
NA
##      new_totbor_24 pair_id treated  bpl
## 58               0      12       0 <NA>
## 2315        249000      30       1 <NA>
## 2393         50000      36       0 <NA>
## 3161             0      44       0 <NA>

# The strength of this dummy is that it helps identify poorest of the poor ho
useholds using a global standard of a poverty line, which is comparable acros
s countries. However, the negative is that income might be misreported and th
e distribution of income within the household might be unequal, as is the cas
e in many developing countries where income of males is often higher than fem
ales and children. If I could ask addditional questions, I would ask question
s about the household consumption and also individual consumption if possible
because a) the reporting is likely to be more accurate b) individual level ef
fects would become more pronounced. I would also ask about the seasonality of
income because the staggered nature of income might cause acute poverty in ce
rtain months, inducing borrowing, which is measured over a longer period, cau
sing disparity in comparison.

# Reading baseline data into R by converting the .dta file to .csv
convert("baseline_controls.dta", "baseline.csv")
baseline <- read.csv("baseline.csv", header = T)

# Understanding the data structure and getting summary stats
str(baseline)

## 'data.frame':    4066 obs. of  17 variables:
##  $ hhid               : int  73 86 179 192 261 268 294 353 500 554 ...
##  $ group_id           : int  35 3 4 76 14 96 14 13 134 122 ...
##  $ hhnomembers        : int  5 4 5 2 7 5 2 5 6 5 ...
##  $ gender_hoh         : int  0 1 1 1 1 1 1 1 0 1 ...
##  $ age_hoh            : int  30 55 51 57 46 48 75 48 60 58 ...
##  $ educyears_hoh      : int  10 10 8 12 19 0 19 16 0 10 ...
##  $ readwrite_hoh      : int  1 1 1 1 1 0 1 1 0 1 ...
##  $ noclasspassed_hoh  : int  0 0 0 0 0 1 0 0 1 0 ...
```

```
##  $ higheduc_hoh       : int  0 0 0 0 1 0 1 1 0 0 ...
##  $ hhnomembers_above18: int  2 4 5 2 4 5 2 4 3 4 ...
##  $ hhnomembers_below18: int  3 0 0 0 3 0 0 1 3 1 ...
##  $ hhreg_muslim       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hhreg_christian    : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ hhcaste_fc         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hhcaste_bc         : int  0 1 0 1 0 0 1 1 1 1 ...
##  $ hhcaste_mbc        : int  1 0 0 0 0 1 0 0 0 0 ...
##  $ hhcaste_sc_st      : int  0 0 1 0 1 0 0 0 0 0 ...
```

```
dim(baseline)
```

```
## [1] 4066   17
```

```
summary(baseline)
```

```
##       hhid           group_id      hhnomembers       gender_hoh
##  Min.   :    73   Min.   :  1.0   Min.   : 1.000   Min.   :0.0000
##  1st Qu.: 76336   1st Qu.: 76.0   1st Qu.: 3.000   1st Qu.:0.0000
##  Median :113767   Median :128.0   Median : 4.000   Median :1.0000
##  Mean   :104647   Mean   :112.5   Mean   : 4.523   Mean   :0.7228
##  3rd Qu.:124972   3rd Qu.:158.0   3rd Qu.: 6.000   3rd Qu.:1.0000
##  Max.   :185460   Max.   :183.0   Max.   :16.000   Max.   :1.0000
##
##     age_hoh      educyears_hoh    readwrite_hoh    noclasspassed_hoh
##  Min.   :19.00   Min.   : 0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:37.00   1st Qu.: 7.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :45.00   Median : 7.000   Median :1.0000   Median :0.0000
##  Mean   :46.68   Mean   : 7.486   Mean   :0.6235   Mean   :0.2265
##  3rd Qu.:56.00   3rd Qu.:11.000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :97.00   Max.   :19.000   Max.   :1.0000   Max.   :1.0000
##
##   higheduc_hoh     hhnomembers_above18 hhnomembers_below18  hhreg_muslim
##  Min.   :0.00000   Min.   : 0.000      Min.   :0.000       Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.: 2.000      1st Qu.:0.000       1st Qu.:0.00000
##  Median :0.00000   Median : 3.000      Median :1.000       Median :0.00000
##  Mean   :0.04796   Mean   : 3.137      Mean   :1.382       Mean   :0.03127
##  3rd Qu.:0.00000   3rd Qu.: 4.000      3rd Qu.:2.000       3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :12.000      Max.   :8.000       Max.   :1.00000
##                                                            NA's   :4
##  hhreg_christian     hhcaste_fc         hhcaste_bc      hhcaste_mbc
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.000000   Median :0.0000   Median :0.0000
##  Mean   :0.04924   Mean   :0.006908   Mean   :0.4041   Mean   :0.3348
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.0000   Max.   :1.0000
##  NA's   :4         NA's   :13         NA's   :13       NA's   :13
##  hhcaste_sc_st
##  Min.   :0.0000
##  1st Qu.:0.0000
```

```
##  Median :0.0000
##  Mean   :0.2539
##  3rd Qu.:1.0000
##  Max.   :1.0000
##  NA's   :13
```

```
# Converting certain variable classes from integer to factor as appropriate a
nd checking data structure again
baseline[,c(1,2,4,7,8,12:17)] <- lapply(baseline[,c(1,2,4,7,8,12:17)], as.fac
tor)
str(baseline)
```

```
## 'data.frame':    4066 obs. of  17 variables:
##  $ hhid               : Factor w/ 4066 levels "73","86","179",..: 1 2 3 4
5 6 7 8 9 10 ...
##  $ group_id           : Factor w/ 101 levels "1","2","3","4",..: 14 3 4 27
6 40 6 5 54 47 ...
##  $ hhnomembers        : int  5 4 5 2 7 5 2 5 6 5 ...
##  $ gender_hoh         : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 1 2 ..
.
##  $ age_hoh            : int  30 55 51 57 46 48 75 48 60 58 ...
##  $ educyears_hoh      : int  10 10 8 12 19 0 19 16 0 10 ...
##  $ readwrite_hoh      : Factor w/ 2 levels "0","1": 2 2 2 2 2 1 2 2 1 2 ..
.
##  $ noclasspassed_hoh  : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 2 1 ..
.
##  $ higheduc_hoh       : int  0 0 0 0 1 0 1 1 0 0 ...
##  $ hhnomembers_above18: int  2 4 5 2 4 5 2 4 3 4 ...
##  $ hhnomembers_below18: int  3 0 0 0 3 0 0 1 3 1 ...
##  $ hhreg_muslim       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ..
.
##  $ hhreg_christian    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ..
.
##  $ hhcaste_fc         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ..
.
##  $ hhcaste_bc         : Factor w/ 2 levels "0","1": 1 2 1 2 1 1 2 2 2 2 ..
.
##  $ hhcaste_mbc        : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 1 1 ..
.
##  $ hhcaste_sc_st      : Factor w/ 2 levels "0","1": 1 1 2 1 2 1 1 1 1 1 ..
.
```

```
# Getting data on households present in both baseline and endline surveys
common_end_base <- intersect(endline_merged$hhid, baseline$hhid)
length(common_end_base)
```

```
## [1] 3802
```

```
# There are 3,802 common households
```

```r
# Identifying which endline households are in the common dataset. w gives a t
rue/false value based on occurrence in the common dataset or not
w <- endline_merged$hhid %in% common_end_base

# Creating another data frame which only contains households from the endline
that are also present in baseline
endline_merged_2 <- data.frame(hhid = endline_merged$hhid[w])

# Checking dimensions and structure of the data frame - still have 3,802 obse
rvations
dim(endline_merged_2)
```

```
## [1] 3802    1
```

```r
head(endline_merged_2)
```

```
##     hhid
## 1 129607
## 2 130298
## 3 130409
## 4 130413
## 5 130299
## 6 130290
```

```r
# Now, merging the new dataset with baseline, retaining both, common househol
ds and those which were present only in baseline, adding both rows and column
s
endline_merged_2 <- merge(endline_merged_2, baseline[, names(baseline)], all
= T)

# Checking dimensions of dataset - now we have 4,066 households, indicating a
n addition of 264 households that were present in baseline but not endline
dim(endline_merged_2)
```

```
## [1] 4066   17
```

```r
# Now combining the data with common and baseline-only households with endlin
e-only households, but retaining only common values for the moment
eb_combined <- merge(endline_merged_2, endline_merged, by = c("hhid", "group_
id", "hhnomembers"))

# Checking dimensions of the fully combined dataset. It seems a little off si
nce the #households dropped from 3,802 to 3,800
dim(eb_combined)
```

```
## [1] 3800   28
```

```r
names(eb_combined)
```

```
##  [1] "hhid"                "group_id"            "hhnomembers"
##  [4] "gender_hoh"          "age_hoh"             "educyears_hoh"
```

```
##  [7] "readwrite_hoh"        "noclasspassed_hoh"    "higheduc_hoh"
## [10] "hhnomembers_above18"  "hhnomembers_below18"  "hhreg_muslim"
## [13] "hhreg_christian"      "hhcaste_fc"           "hhcaste_bc"
## [16] "hhcaste_mbc"          "hhcaste_sc_st"        "totformalborrow_24"
## [19] "totinformalborrow_24" "hhinc"                "survey_round"
## [22] "new_totformbor_24"    "newtotinformbor_24"   "new_hhinc"
## [25] "new_totbor_24"        "pair_id"              "treated"
## [28] "bpl"
```

```
# Trying to identify why 2 HHs dropped out -- seems like the group_id coding
differs in endline and baseline for these 2 HHs
eb_combined <- merge(endline_merged_2, endline_merged, by = c("hhid", "hhnome
mbers"))

# Identifying and browsing the particular HHs that got dropped. We see that t
he group_id in baseline in 152 while in endline is 148.
eb_combined[eb_combined$group_id.x != eb_combined$group_id.y,]
```

```
##       hhid hhnomembers group_id.x gender_hoh age_hoh educyears_hoh
## 329 106131           3        152          1      49             9
## 353 106360           2        152          1      64             0
##     readwrite_hoh noclasspassed_hoh higheduc_hoh hhnomembers_above18
## 329             1                 0            0                   3
## 353             0                 1            0                   2
##     hhnomembers_below18 hhreg_muslim hhreg_christian hhcaste_fc hhcaste_bc
## 329                   0            0               0          0          1
## 353                   0            0               0          0          1
##     hhcaste_mbc hhcaste_sc_st group_id.y totformalborrow_24
## 329           0             0        148               None
## 353           0             0        148               None
##     totinformalborrow_24 hhinc survey_round new_totformbor_24
## 329               205000  3300   Endline II                 0
## 353                 None  6000   Endline II                 0
##     newtotinformbor_24 new_hhinc new_totbor_24 pair_id treated bpl
## 329             205000      3300        205000      39       0   0
## 353                  0      6000             0      39       0   0
```

```
table(baseline$group_id)
```

```
##
##   1   2   3   4  13  14  16  21  22  28  30  32  34  35  37  38  40  43  4
4  48
##  33  34  29  28  30  38  45  42  40  40  42  45  44  39  42  42  42  40  4
4  41
##  57  58  62  63  64  73  76  77  80  82  83  84  85  86  87  89  91  92  9
4  96
##  41  44  35  41  36  35  35  42  44  42  42  42  45  42  50  39  43  43  4
8  39
##  98 101 103 108 116 120 122 123 124 126 127 128 133 134 135 137 138 139 14
1 142
##  31  39  38  43  41  39  41  45  35  25  41  39  29  44  45  42  38  46  2
```

```
9  43
## 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 16
1 162
##  42  42  44  45  44  40  41  38  41  45  51  44  41  40  40  39  40  42  3
7  40
## 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 18
1 182
##  42  43  45  41  40  41  38  36  37  40  40  44  45  40  40  45  41  33  3
5  40
## 183
##  43
```

```
# It is clear that there is a coding error in one of the surveys. Assuming th
at the baseline categorization was correct, replacing the mis-categorization
in endline with baseline values.

endline_merged[endline_merged$hhid == 106131, 1] <- 152
endline_merged[endline_merged$hhid == 106360, 1] <- 152

# Now, creating the final merged dataset with endline-only, baseline-only, an
d common households
eb_combined <- merge(endline_merged_2, endline_merged, by = c("hhid", "group_
id", "hhnomembers"), all = T)

# Checking row dimensions: sense-check the answer, which should be 4,066 + 4,
160 - 3802 = 4,424. # Checking if all columns have been included. Again, sens
e-check: answer should be 14 + 17 - 3 = 28. The answer is verified. Therefore
, the data merging process should have worked fine.
dim(eb_combined)
```

```
## [1] 4424   28
```

```
# Checking the data structure and vital signs.
names(eb_combined)
```

```
##  [1] "hhid"                 "group_id"             "hhnomembers"
##  [4] "gender_hoh"           "age_hoh"              "educyears_hoh"
##  [7] "readwrite_hoh"        "noclasspassed_hoh"    "higheduc_hoh"
## [10] "hhnomembers_above18"  "hhnomembers_below18"  "hhreg_muslim"
## [13] "hhreg_christian"      "hhcaste_fc"           "hhcaste_bc"
## [16] "hhcaste_mbc"          "hhcaste_sc_st"        "totformalborrow_24"
## [19] "totinformalborrow_24" "hhinc"                "survey_round"
## [22] "new_totformbor_24"    "newtotinformbor_24"   "new_hhinc"
## [25] "new_totbor_24"        "pair_id"              "treated"
## [28] "bpl"
```

```
summary(eb_combined)
```

```
##      hhid              group_id     hhnomembers     gender_hoh     age_hoh
##  Length:4424        153    :  57   Min.   : 1.000   0   :1127   Min.   :19
.00
```

```
##  Class :character   82     :  52   1st Qu.: 3.000   1   :2939   1st Qu.:37
.00
##  Mode  :character   87     :  52   Median : 4.000   NA's: 358   Median :45
.00
##                     135    :  52   Mean   : 4.499               Mean   :46
.68
##                     139    :  51   3rd Qu.: 5.000               3rd Qu.:56
.00
##                     147    :  51   Max.   :16.000               Max.   :97
.00
##                     (Other):4109                                NA's   :35
8
##  educyears_hoh    readwrite_hoh noclasspassed_hoh  higheduc_hoh
##  Min.   : 0.000   0   :1531     0   :3145         Min.   :0.000
##  1st Qu.: 7.000   1   :2535     1   : 921         1st Qu.:0.000
##  Median : 7.000   NA's: 358     NA's: 358         Median :0.000
##  Mean   : 7.487                                   Mean   :0.048
##  3rd Qu.:11.000                                   3rd Qu.:0.000
##  Max.   :19.000                                   Max.   :1.000
##  NA's   :358                                      NA's   :358
##  hhnomembers_above18 hhnomembers_below18 hhreg_muslim hhreg_christian
##  Min.   : 0.000      Min.   :0.000       0   :3935    0   :3862
##  1st Qu.: 2.000      1st Qu.:0.000       1   : 127    1   : 200
##  Median : 3.000      Median :1.000       NA's: 362    NA's: 362
##  Mean   : 3.137      Mean   :1.382
##  3rd Qu.: 4.000      3rd Qu.:2.000
##  Max.   :12.000      Max.   :8.000
##  NA's   :358         NA's   :358
##  hhcaste_fc  hhcaste_bc  hhcaste_mbc hhcaste_sc_st totformalborrow_24
##  0   :4025   0   :2415   0   :2696   0   :3024     None   :1221
##  1   :  28   1   :1638   1   :1357   1   :1029     20000  : 242
##  NA's: 371   NA's: 371   NA's: 371   NA's: 371     50000  : 133
##                                                   30000  : 128
##                                                   25000  : 120
##                                                   (Other):2316
##                                                   NA's   : 264
##  totinformalborrow_24     hhinc           survey_round  new_totformbor_24
##  None   :1539         None   : 244   Endline I  : 641   Min.   :     0
##  10000  : 212         10000  : 162   Endline II :2873   1st Qu.:     0
##  5000   : 134         3000   : 156   Endline III: 646   Median : 30000
##  50000  : 117         1000   : 135   NA's       : 264   Mean   : 59758
##  20000  : 115         5000   : 133                      3rd Qu.: 75000
##  (Other):2043         (Other):3330                      Max.   :446861
##  NA's   : 264         NA's   : 264                      NA's   :268
##  newtotinformbor_24   new_hhinc      new_totbor_24       pair_id
##  Min.   :     0     Min.   :     0   Min.   :     0   40     : 122
##  1st Qu.:     0     1st Qu.:  2850   1st Qu.: 20000   39     : 102
##  Median : 10000     Median :  6000   Median : 56000   18     :  96
##  Mean   : 37426     Mean   : 10450   Mean   : 97090   38     :  95
##  3rd Qu.: 45000     3rd Qu.: 11000   3rd Qu.:126175   23     :  94
```

```
##  Max.   :295350     Max.   :214190   Max.   :742211   (Other):3651
##  NA's   :268        NA's   :268      NA's   :264      NA's   : 264
##  treated       bpl
##  0   :2048   0   :2914
##  1   :2112   1   :1242
##  NA's: 264   NA's: 268
##
##
##
##
```

```
str(eb_combined)
```

```
## 'data.frame':    4424 obs. of  28 variables:
##  $ hhid                : chr  "100003" "100005" "100039" "100068" ...
##  $ group_id            : Factor w/ 101 levels "1","2","3","4",..: 67 67 68
68 68 68 67 67 67 67 ...
##  $ hhnomembers         : int  4 3 4 4 3 7 5 1 2 6 ...
##  $ gender_hoh          : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 1 2 2 .
..
##  $ age_hoh             : int  45 70 45 39 48 80 57 70 88 44 ...
##  $ educyears_hoh       : int  12 0 0 11 0 0 7 0 12 10 ...
##  $ readwrite_hoh       : Factor w/ 2 levels "0","1": 2 1 1 2 1 1 2 1 2 1 .
..
##  $ noclasspassed_hoh   : Factor w/ 2 levels "0","1": 1 2 2 1 2 2 1 2 1 1 .
..
##  $ higheduc_hoh        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hhnomembers_above18 : int  3 3 3 2 3 4 4 1 2 3 ...
##  $ hhnomembers_below18 : int  1 0 1 2 0 3 1 0 0 3 ...
##  $ hhreg_muslim        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 .
..
##  $ hhreg_christian     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 .
..
##  $ hhcaste_fc          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 .
..
##  $ hhcaste_bc          : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 1 1 1 1 .
..
##  $ hhcaste_mbc         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 .
..
##  $ hhcaste_sc_st       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 2 2 2 .
..
##  $ totformalborrow_24  : Factor w/ 414 levels ".","10000","100000",..: 225
414 130 204 84 73 349 414 414 343 ...
##  $ totinformalborrow_24: Factor w/ 347 levels ".","10000","100000",..: 264
167 318 24 167 268 206 149 347 222 ...
##  $ hhinc               : Factor w/ 803 levels "100","1000","10000",..: 681
2 112 561 3 722 67 161 2 525 ...
##  $ survey_round        : Factor w/ 3 levels "Endline I","Endline II",..: 2
2 2 2 2 2 2 2 2 2 ...
##  $ new_totformbor_24   : num  33000 0 200000 290000 16000 150000 68000 0 0
```

```
65000 ...
##  $ newtotinformbor_24  : num  56500 28000 81000 112000 28000 58000 35000 2
500 0 40000 ...
##  $ new_hhinc           : num  7200 1000 12800 5000 10000 8000 11500 1500 1
000 4500 ...
##  $ new_totbor_24       : num  89500 28000 281000 402000 44000 208000 10300
0 2500 0 105000 ...
##  $ pair_id             : Factor w/ 50 levels "1","2","4","5",..: 35 35 35
35 35 35 35 35 35 35 ...
##  $ treated             : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 2 2 2 2 .
..
##  $ bpl                 : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 2 2 .
..
```

```
# To deal with baseline-only and endline-only households, we create a dummy t
o identify which is which. This dummy takes on the value 0 if common between
baseline and endline, 1 if baseline-only, and 2 if endline-only.
eb_combined$missing_status <- ifelse(eb_combined$hhid %in% common_end_base, 0
, ifelse(eb_combined$hhid %in% baseline$hhid, 1, 2))

# Checking how many values we end up with for each dummy - the answer ties in
with the answers we have got previously so the process works.
table(eb_combined$missing_status)
```

```
##
##    0    1    2
## 3802  264  358
```

```
# I have added dummy variables instead of dropping observations because it is
important to not mess up the balance between the treatment and control groups
. Dummies allow one to analyze the difference, if any, between the characteri
stics of the three categories of households, i.e. common, endline-only, basel
ine-only.
```

Data Analysis

```
# The testable hyotheses could be whether a) access to formal credit reduces
informal lending and increases formal lending (expect formal borrowing to inc
rease and informal borrowing to decrease) b) access to more/better credit ter
ms increases household income (expect hh income to increase) c) savings respo
nd to better credit terms (expect savings to increase)

# Choosing the following variables to test because they are expected to have
an impact on key outcome variables and hence we want to make sure that the gr
oups are 'balanced', i.e. they are not statistically significantly different
from each other. This can be seen by the p-values of the following t-tests, a
ll of which are >0.05, so we fail to reject that the two groups are significa
ntly different from each other. This means our randomization is valid and so
is our experiment and its conclusions.

# Demographics
```

```r
t1 <- t.test(as.numeric(eb_combined$hhid) ~ eb_combined$treated)
t2 <- t.test(as.numeric(eb_combined$hhcaste_sc_st) ~ eb_combined$treated)
t3 <- t.test(as.numeric(eb_combined$hhcaste_fc) ~ eb_combined$treated)

# Income
t4 <- t.test(as.numeric(eb_combined$new_hhinc) ~ eb_combined$treated)
t5 <- t.test(as.numeric(eb_combined$bpl) ~ eb_combined$treated)

# Characteristics of the head of household
t6 <- t.test(as.numeric(eb_combined$gender_hoh) ~ eb_combined$treated)
t7 <- t.test(as.numeric(eb_combined$age_hoh) ~ eb_combined$treated)
t8 <- t.test(as.numeric(eb_combined$educyears_hoh) ~ eb_combined$treated)
t9 <- t.test(as.numeric(eb_combined$readwrite_hoh) ~ eb_combined$treated)
t10 <- t.test(as.numeric(eb_combined$noclasspassed_hoh) ~ eb_combined$treated
)

t <- list(t1,t2,t3,t4,t5,t6,t7,t8,t9,t10)

t
```

```
## [[1]]
##
##  Welch Two Sample t-test
##
## data:  as.numeric(eb_combined$hhid) by eb_combined$treated
## t = -1.3431, df = 4149.3, p-value = 0.1793
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4828.9407   902.5691
## sample estimates:
## mean in group 0 mean in group 1
##        104523.4        106486.6
##
##
## [[2]]
##
##  Welch Two Sample t-test
##
## data:  as.numeric(eb_combined$hhcaste_sc_st) by eb_combined$treated
## t = 0.25298, df = 3793.9, p-value = 0.8003
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02429913  0.03149885
## sample estimates:
## mean in group 0 mean in group 1
##        1.261497        1.257897
##
##
## [[3]]
##
```

```
##  Welch Two Sample t-test
##
## data:  as.numeric(eb_combined$hhcaste_fc) by eb_combined$treated
## t = -1.1541, df = 3649.8, p-value = 0.2485
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.007975366  0.002065043
## sample estimates:
## mean in group 0 mean in group 1
##        1.004813        1.007768
##
##
## [[4]]
##
##  Welch Two Sample t-test
##
## data:  as.numeric(eb_combined$new_hhinc) by eb_combined$treated
## t = -1.2218, df = 4153.4, p-value = 0.2218
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1764.0338   409.4706
## sample estimates:
## mean in group 0 mean in group 1
##        10106.28        10783.56
##
##
## [[5]]
##
##  Welch Two Sample t-test
##
## data:  as.numeric(eb_combined$bpl) by eb_combined$treated
## t = 1.2103, df = 4144.3, p-value = 0.2262
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01065864  0.04505015
## sample estimates:
## mean in group 0 mean in group 1
##        1.307579        1.290384
##
##
## [[6]]
##
##  Welch Two Sample t-test
##
## data:  as.numeric(eb_combined$gender_hoh) by eb_combined$treated
## t = 1.0822, df = 3799.3, p-value = 0.2793
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01267773  0.04391357
## sample estimates:
```

```
## mean in group 0 mean in group 1
##        1.735970        1.720352
##
##
## [[7]]
##
##  Welch Two Sample t-test
##
## data:  as.numeric(eb_combined$age_hoh) by eb_combined$treated
## t = -1.5165, df = 3799.9, p-value = 0.1295
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.4159939  0.1808706
## sample estimates:
## mean in group 0 mean in group 1
##        46.44148        47.05904
##
##
## [[8]]
##
##  Welch Two Sample t-test
##
## data:  as.numeric(eb_combined$educyears_hoh) by eb_combined$treated
## t = 0.013561, df = 3798.1, p-value = 0.9892
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3051242  0.3093746
## sample estimates:
## mean in group 0 mean in group 1
##        7.435061        7.432936
##
##
## [[9]]
##
##  Welch Two Sample t-test
##
## data:  as.numeric(eb_combined$readwrite_hoh) by eb_combined$treated
## t = -0.22977, df = 3795.8, p-value = 0.8183
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03452292  0.02727999
## sample estimates:
## mean in group 0 mean in group 1
##        1.616782        1.620404
##
##
## [[10]]
##
##  Welch Two Sample t-test
##
```

```
## data:  as.numeric(eb_combined$noclasspassed_hoh) by eb_combined$treated
## t = -1.2045, df = 3799.8, p-value = 0.2285
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04316967  0.01031241
## sample estimates:
## mean in group 0 mean in group 1
##        1.221272        1.237701
```

```
# The estimates I got here is the best that I was able to do in making a bala
nce table of t-tests - it's not ideal but given the limited time, I could not
think of anything else other than the following method.

# Tried making a loop but not getting too far - sample code:
#t_tests <- data.frame(rep(rep(NA,10),10))
#test_vars <- data.frame(eb_combined$hhid, eb_combined$hhcaste_sc_st, eb_comb
ined$hhcaste_fc, eb_combined$new_hhinc, eb_combined$bpl, #eb_combined$gender_
hoh, eb_combined$age_hoh, eb_combined$educyears_hoh, eb_combined$readwrite_ho
h, eb_combined$noclasspassed_hoh)

#for (i in c(1:10)) {
#  t_tests[[i]] <- t.test(as.numeric(test_vars[[i]]) ~ eb_combined$treated)
#}

#t_tests[[1]]
```

```
library("clubSandwich") # helps test for coefficients by clustering standard
errors
```

```
## Warning: package 'clubSandwich' was built under R version 3.6.3
```

```
## Registered S3 method overwritten by 'clubSandwich':
##   method    from
##   bread.mlm sandwich
```

```
library("plm") # helps run  fixed effects linear model
```

```
## Warning: package 'plm' was built under R version 3.6.3
```

```
# Running OLS regressing household income on treatment dummy, with pair fixed
effects
hh_inc_on_treatment <- lm(eb_combined$new_hhinc ~ eb_combined$treated + eb_co
mbined$pair_id - 1)
summary(hh_inc_on_treatment)
```

```
##
## Call:
## lm(formula = eb_combined$new_hhinc ~ eb_combined$treated + eb_combined$pai
r_id -
##     1)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -24512  -7106  -3634   1558 204831
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## eb_combined$treated0   11588.9     1983.3   5.843 5.52e-09 ***
## eb_combined$treated1   12303.9     1980.5   6.213 5.73e-10 ***
## eb_combined$pair_id2   -1233.0     2767.2  -0.446  0.65591
## eb_combined$pair_id4   -1711.6     2767.2  -0.619  0.53625
## eb_combined$pair_id5   -5199.4     2784.3  -1.867  0.06192 .
## eb_combined$pair_id6   -1638.2     2735.0  -0.599  0.54922
## eb_combined$pair_id7     254.9     2830.6   0.090  0.92824
## eb_combined$pair_id9   -2129.0     2802.2  -0.760  0.44744
## eb_combined$pair_id10  -3423.6     2719.9  -1.259  0.20820
## eb_combined$pair_id11  -4820.1     2742.8  -1.757  0.07893 .
## eb_combined$pair_id12  -4200.8     2861.1  -1.468  0.14211
## eb_combined$pair_id13  -2882.0     2802.4  -1.028  0.30381
## eb_combined$pair_id14  -5198.2     2758.9  -1.884  0.05961 .
## eb_combined$pair_id15   3507.9     2821.7   1.243  0.21388
## eb_combined$pair_id16  -4482.9     2882.9  -1.555  0.12003
## eb_combined$pair_id17  -4256.7     2871.9  -1.482  0.13836
## eb_combined$pair_id18  -2944.7     2665.0  -1.105  0.26925
## eb_combined$pair_id20   -509.3     2750.8  -0.185  0.85313
## eb_combined$pair_id21  -4239.6     2758.9  -1.537  0.12444
## eb_combined$pair_id22  -3992.5     2882.6  -1.385  0.16612
## eb_combined$pair_id23    304.7     2678.0   0.114  0.90943
## eb_combined$pair_id25  -4829.3     2995.7  -1.612  0.10702
## eb_combined$pair_id26  -3228.8     2750.8  -1.174  0.24056
## eb_combined$pair_id28  -5190.7     2705.4  -1.919  0.05510 .
## eb_combined$pair_id29  -4541.7     2784.3  -1.631  0.10293
## eb_combined$pair_id30   2515.7     2821.0   0.892  0.37257
## eb_combined$pair_id31  -1971.4     3008.7  -0.655  0.51236
## eb_combined$pair_id32   -327.3     2750.8  -0.119  0.90530
## eb_combined$pair_id33  -3727.0     2750.7  -1.355  0.17552
## eb_combined$pair_id34   -442.6     2784.4  -0.159  0.87370
## eb_combined$pair_id35  -3655.6     2830.8  -1.291  0.19664
## eb_combined$pair_id36  -6015.3     2705.4  -2.223  0.02624 *
## eb_combined$pair_id37  -2093.4     2678.0  -0.782  0.43444
## eb_combined$pair_id38  -1862.4     2671.4  -0.697  0.48575
## eb_combined$pair_id39  -2161.4     2628.9  -0.822  0.41103
## eb_combined$pair_id40  -1855.8     2533.7  -0.732  0.46393
## eb_combined$pair_id41   1638.0     2712.6   0.604  0.54597
## eb_combined$pair_id42   1470.9     2705.4   0.544  0.58668
## eb_combined$pair_id43    337.5     2742.8   0.123  0.90208
## eb_combined$pair_id44   -189.9     2784.4  -0.068  0.94563
## eb_combined$pair_id45  12207.9     2784.3   4.385 1.19e-05 ***
## eb_combined$pair_id46   7571.7     2712.5   2.791  0.00527 **
## eb_combined$pair_id47  -4834.9     2742.8  -1.763  0.07801 .
## eb_combined$pair_id48   4416.4     2742.8   1.610  0.10743
```

```
## eb_combined$pair_id49    341.8     2811.5   0.122  0.90325
## eb_combined$pair_id50  -3779.9     2742.8  -1.378  0.16824
## eb_combined$pair_id51   1448.4     2698.3   0.537  0.59146
## eb_combined$pair_id52  -2203.0     2742.8  -0.803  0.42191
## eb_combined$pair_id53   -971.8     2727.4  -0.356  0.72164
## eb_combined$pair_id54  -2276.1     2775.7  -0.820  0.41225
## eb_combined$pair_id55  -3954.8     2820.9  -1.402  0.16100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17660 on 4105 degrees of freedom
##   (268 observations deleted due to missingness)
## Multiple R-squared:  0.2812, Adjusted R-squared:  0.2723
## F-statistic: 31.49 on 51 and 4105 DF,  p-value: < 2.2e-16
```

```
plm_model <- plm(new_hhinc ~ treated, data = eb_combined[!is.na(eb_combined$n
ew_hhinc),], index = c("pair_id"), model = "within")
summary(plm_model)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = new_hhinc ~ treated, data = eb_combined[!is.na(eb_combined$n
ew_hhinc),
##     ], model = "within", index = c("pair_id"))
##
## Unbalanced Panel: n = 50, T = 60-122, N = 4156
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -24511.7  -7106.2  -3634.1   1558.5 204831.0
##
## Coefficients:
##          Estimate Std. Error t-value Pr(>|t|)
## treated1   714.99     550.52  1.2987   0.1941
##
## Total Sum of Squares:    1.2814e+12
## Residual Sum of Squares: 1.2808e+12
## R-Squared:      0.00041073
## Adj. R-Squared: -0.011765
## F-statistic: 1.68674 on 1 and 4105 DF, p-value: 0.1941
```

```
# It is appropriate to use a fixed effects specification here because we want
to control for time-invariant characteristics at the pair level and isolate t
he effects of the treatment. In this case, at a pair level, except for one pa
ir, the pair fixed effects are statistically significant at 0.1% level. This
provides validity to our fixed effects specification, suggesting that pair-le
vel characteristics do explain the variation in HH income.

# The point estimate is the difference between means of the treatment and con
```

```
trol groups as given by the plm_model, which shows that the difference in HH
income between treatment and control groups is Rs. 715 and is not statistical
ly different from zero, i.e. we fail to reject that the treatment caused a si
gnificant increase in HH income for the treated HHs.

# Testing coefficient after clustering standard errors at the group level ->
corrected standard errors
coef_test(plm_model, vcov = "CR2", cluster = eb_combined$group_id, test = "Sa
tterthwaite")

##      Coef. Estimate  SE t-stat d.f. p-val (Satt) Sig.
## 1 treated1      715 541   1.32 96.5        0.189

# It is reasonable for us to cluster standard errors at the group_id level be
cause it represents a certain area for service delivery and we expect errors
to be correlated within those areas. Even after correcting for standard error
s, the treatment effect is not statistically significant.

#Redefining a log(hhinc) variable
eb_combined$new_log_hhinc <- log(eb_combined$new_hhinc)

# Defining new data for which log(hhinc) is not NA or -Inf
new_data <- eb_combined[!is.na(eb_combined$new_log_hhinc) & eb_combined$new_l
og_hhinc > 0,]

#Checking dimensions of the new data
dim(new_data)

## [1] 3912   30

# Running a log specification with pair fixed effects :
log_hh_inc_on_treatment <- plm(new_log_hhinc ~ treated, data = new_data, inde
x = "pair_id", model = "within")

summary(log_hh_inc_on_treatment)

## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = new_log_hhinc ~ treated, data = new_data, model = "within",
##     index = "pair_id")
##
## Unbalanced Panel: n = 50, T = 50-117, N = 3912
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -4.991951 -0.630343  0.069142  0.636216  3.672067
##
## Coefficients:
##          Estimate Std. Error t-value Pr(>|t|)
## treated1 0.062811   0.033636  1.8674  0.06192 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    4234.9
## Residual Sum of Squares: 4231.1
## R-Squared:      0.00090234
## Adj. R-Squared: -0.012036
## F-statistic: 3.48709 on 1 and 3861 DF, p-value: 0.061925
```

```
# Running a log specification leads our coeeficient, that is treatment effect to be significant at 10% level. It brings down the standard error comparitively and at 10% level, we can reject that there was no increase in HH income due to the treatment. Since we have a smaller set of observations here, we are compromising a bit on the power of our test.

# Re-running the previous specification with household-level controls (age, gender, education, caste, religion, members over 18 years of age)
log_hh_inc_on_treatment_controls <- plm(new_log_hhinc ~ treated + age_hoh + gender_hoh + educyears_hoh + hhcaste_sc_st + hhcaste_fc + hhreg_muslim + hhnomembers_above18 + hhnomembers, data = new_data, index = "pair_id", model = "within")

summary(log_hh_inc_on_treatment_controls)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = new_log_hhinc ~ treated + age_hoh + gender_hoh +
##     educyears_hoh + hhcaste_sc_st + hhcaste_fc + hhreg_muslim +
##     hhnomembers_above18 + hhnomembers, data = new_data, model = "within",
##     index = "pair_id")
##
## Unbalanced Panel: n = 50, T = 46-111, N = 3581
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -5.03761 -0.57939  0.05752  0.60146  3.92676
##
## Coefficients:
##                       Estimate Std. Error t-value  Pr(>|t|)
## treated1             0.0757404  0.0333379  2.2719 0.0231528 *
## age_hoh             -0.0068194  0.0015374 -4.4358 9.453e-06 ***
## gender_hoh1          0.1687348  0.0411687  4.0986 4.250e-05 ***
## educyears_hoh        0.0311659  0.0038208  8.1569 4.733e-16 ***
## hhcaste_sc_st1      -0.1280467  0.0397060 -3.2249 0.0012718 **
## hhcaste_fc1          0.3264034  0.2157967  1.5126 0.1304836
## hhreg_muslim1        0.1309620  0.1074364  1.2190 0.2229364
## hhnomembers_above18  0.1103127  0.0196074  5.6261 1.987e-08 ***
## hhnomembers          0.0490860  0.0144201  3.4040 0.0006715 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    3830.4
## Residual Sum of Squares: 3448.8
## R-Squared:      0.099634
## Adj. R-Squared: 0.084807
## F-statistic: 43.3046 on 9 and 3522 DF, p-value: < 2.22e-16
```

```
# I chose these controls because they are intuitively likely to explain varia
tion in hh income; the fact that almost all are significant shows that the tr
eatment effect could have been biased earlier, due to the omitted variables b
ias, since we failed to account for key factors that are correlated with both
the treatment and the hh income. After including hh level controls, we find t
hat our treatment effects becomes significant at 5% level as the magnitude of
our estimate increases.

#Creating publication quality regression output in LaTeX

library("stargazer")

stargazer(log_hh_inc_on_treatment_controls, title = "Regression Results with
Household Level Controls", dep.var.labels = c("Log of Household Income over l
ast 30 days"), covariate.labels = c("Treatment", "Age (Head of Household)", "
Gender (Head of Household)", "Years of Education (Head of Household)", "Caste
- SC/ST", "Caste - Forward", "Religion - Muslim", "No. of Household members o
ver age of 18", "No. of Household members"))

#Defining income quartiles
quantile(eb_combined$new_hhinc, c(0.25, 0.5, 0.75, 1), na.rm = T)
```

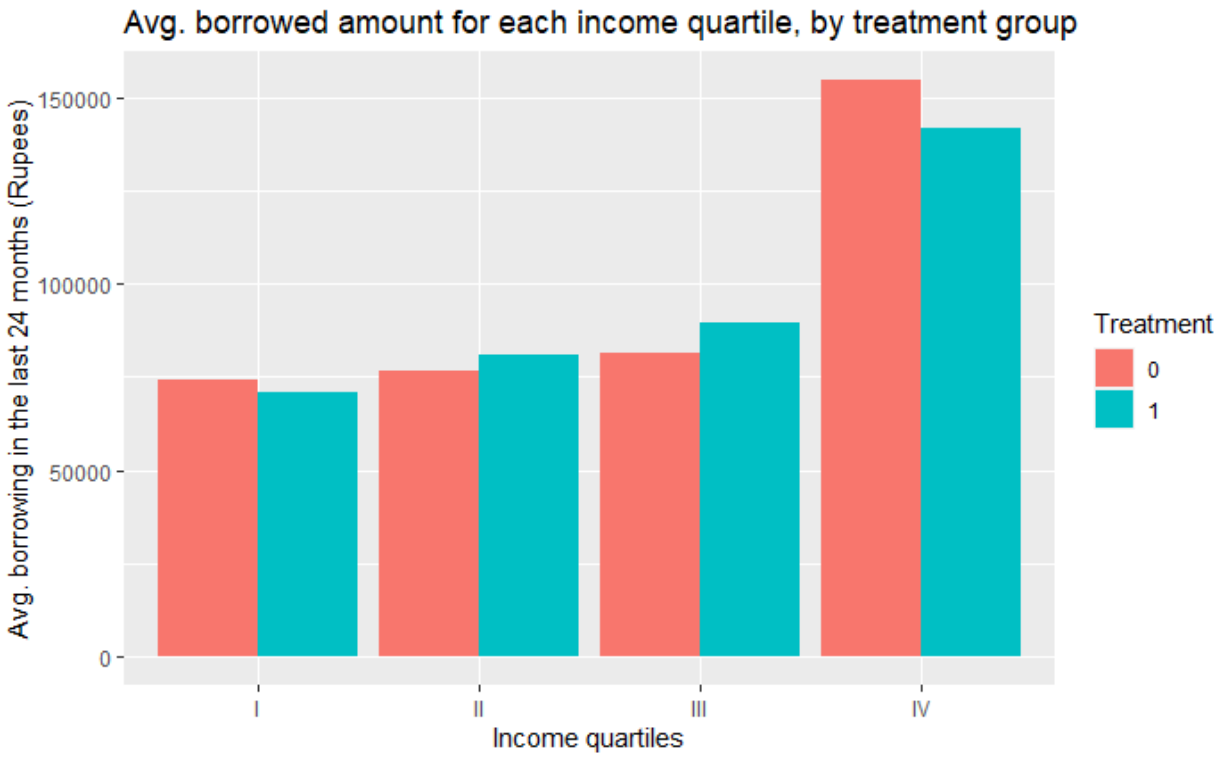
```
##      25%      50%      75%     100%
##   2850.0   6000.0  11000.0 214190.3
```

```
eb_combined$new_hhinc_quartile <- ifelse(eb_combined$new_hhinc < 2850, "I", i
felse(eb_combined$new_hhinc < 6000, "II", ifelse(eb_combined$new_hhinc < 1100
0, "III", ifelse(eb_combined$new_hhinc <= 214190.3, "IV", NA))))

# Creating a data frame to plot the barchart
avg_borr_inc <- aggregate(eb_combined$new_totbor_24, by = list(eb_combined$tr
eated, eb_combined$new_hhinc_quartile), FUN = mean)
avg_borr_inc <- data.frame(Treatment = avg_borr_inc[[1]], IncomeQuartile = av
g_borr_inc[[2]], AvgBorrowing = avg_borr_inc[[3]])

# Plotting barchart
library("ggplot2")

ggplot(avg_borr_inc, aes(IncomeQuartile, AvgBorrowing)) + geom_bar(aes(fill =
Treatment), stat = "identity", position = "dodge") + labs(x = "Income quartil
es", y = "Avg. borrowing in the last 24 months (Rupees)", title = "Avg. borro
wed amount for each income quartile, by treatment group")
```


Avg. borrowed amount for each income quartile, by treatment group
Avg. borrowing in the last 24 months (Rupees)
150000
100000
50000
0
I
II
III
IV
Income quartiles
Treatment
0
1