Problem Set 4: Randomized Control Trials

Claire Duquennois

**Name:** Xiang Li

Instructions:

* This assignment is an individual assignment. You may discuss your responses in small groups or reach out to a classmate if you are having difficulties with coding but your code and answers must reflect your individual efforts and be written using your own words. Identical assignments will be given a zero grade.

# 1. Empirical Analysis using Data from Bryan, G., Chowdury, S., Mobarak, A. M. (2014, Econometrica)

This exercise uses data from Bryan,Chowdhury, and Mobarak’s paper, “Underinvestment in a Profitable Technology: the Case of Seasonal Migration in Bangladesh,” published in *Econometrica* in 2014. This paper studies the effects of seasonal migration on household consumption during the lean season in rural Bangladesh by randomly subsidizing the cost of seasonal migration.

# 2. Set Up:

## 2.1 Finding the data

The data can be found by going to Mushfiq Mobarak’s Yale faculty page, select “data”, and then following the link to the data repository page on the Harvard dataverse. You will need to sign in to get access to the data files. Once logged in, you will find many possible files to download. Navigate to the second page of listed files and download Mobarak - Monga Dataverse files.zip which contains all the files we need.

## 2.2 Question: Loading the data - Load any packages you will need and the data contained in the following files Round1\_Controls\_Table1.dta and Round2.dta. How many observations are contained in each of these datasets. What is the level of an observation? Explain any discrepancies between the datasets.\*\*

**Code and Answer:**

library("haven")  
library("dplyr")

##   
## 载入程辑包：'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library("stargazer")

##   
## Please cite as:

## Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

library("lfe")

## 载入需要的程辑包：Matrix

## Warning: 程辑包'Matrix'是用R版本4.3.2 来建造的

round1 <- read\_dta("Round1\_Controls\_Table1.dta")  
round2 <- read\_dta("Round2.dta")  
  
nrow(round1)

## [1] 1900

nrow(round2)

## [1] 1907

Answer: There are 1900 observations contained in round1 and 1907 observations contained in round2. It’s village/household level of observation. People’s moving made the difference.

## 2.3 Question: (2 pages) Data Description- The dataset contains many variables, some of which are not used in this exercise. Keep the following variables in the final datasets (Hint: use the select function in dplyr).\*\*

For Round 1 data:

| Name | Description |
| --- | --- |
| cash | In cash treatment group |
| credit | In credit treatment group |
| info | In information treatment group |
| control | In control group |
| q9pdcalq9 | Total calories per person per day |
| exp\_total\_pc\_r1 | Total monthly household expenditures per capita |
| hhmembers\_r1 | Number of household members |
| tsaving\_hh\_r1 | Total household savings |
| hhh\_education | Household head is educated |
| num\_adltmalesr1 | Number of adult males in the household |

For Round 2 data:

| Name | Description |
| --- | --- |
| cash | In cash treatment group |
| credit | In credit treatment group |
| info | In information treatment group |
| control | In control group |
| average\_exp2 | Total consumption per person per month in round 2 |
| lit | Highest reading and writing ability of household |
| walls\_good | Wall material (income proxy) |
| monga | Subjective expectations about monga at baseline |
| dhaka\_remit | Subjective expectations about migration remitances at baseline |
| dhaka\_network | Subjective expectations about social network in city at baseline |
| exp\_total\_pc\_r1 | Total household expenditures per capita at baseline |
| subsistencer1 | Share of food out of total expenditures at baseline |
| num\_adltmalesr1 | Household adult males at baseline |
| num\_childrenr1 | Household small children at baseline |
| avgQ13earned | Average skill score of network |
| constrainedr1 | Denied or ineligible for credit at baseline |
| bankedr1 | Has received credit at baseline |
| upazila | Sub-district name |
| village | Village name |
| migrant | Member of household migrates this season |
| total\_fish | Total monthly household expenditures per capita on fish |
| migrant\_new | Household has a first time migrant this season |

**A description of each variable should appear in the column headers of the loaded data.**

**Code:**

selectround1 <- select(round1, cash, credit, info, control, q9pdcalq9, exp\_total\_pc\_r1, hhmembers\_r1, tsaving\_hh\_r1, hhh\_education, num\_adltmalesr1)  
selectround2 <- select(round2, cash, credit, info, control, average\_exp2, lit, walls\_good, monga, dhaka\_remit, dhaka\_network, exp\_total\_pc\_r1, subsistencer1, num\_adltmalesr1, num\_childrenr1, avgQ13earned, constrainedr1, bankedr1, upazila, village, migrant, total\_fish, migrant\_new)

# 3. Analysis:

## 3.1 **Question: Regress all the baseline household characteristics still included in the round 1 data on the following three variables: , and , and present your results in a table. What is the equivalent table in the paper?**

**Code:**

colnames(selectround1)[colnames(selectround1) == "q9pdcalq9"] <- "TotalCalories"  
colnames(selectround1)[colnames(selectround1) == "exp\_total\_pc\_r1"] <- "TotalMonthlyHouseholdExpenditures"  
colnames(selectround1)[colnames(selectround1) == "hhmembers\_r1"] <- "HouseholdMembers"  
colnames(selectround1)[colnames(selectround1) == "tsaving\_hh\_r1"] <- "TotalHouseholdSavings"  
colnames(selectround1)[colnames(selectround1) == "hhh\_education"] <- "HouseholdHeadIsEducated"  
colnames(selectround1)[colnames(selectround1) == "num\_adltmalesr1"] <- "AdultMales\_inthehousehold"

rd1.1 <- felm(TotalCalories ~ cash + credit + info, selectround1)  
rd1.2 <- felm(TotalMonthlyHouseholdExpenditures ~ cash + credit + info, selectround1)  
rd1.3 <- felm(HouseholdMembers ~ cash + credit + info, selectround1)  
rd1.4 <- felm(TotalHouseholdSavings ~ cash + credit + info, selectround1)  
rd1.5 <- felm(HouseholdHeadIsEducated ~ cash + credit + info, selectround1)  
rd1.6 <- felm(AdultMales\_inthehousehold ~ cash + credit + info, selectround1)  
  
stargazer(rd1.1, rd1.2, rd1.3, rd1.4, rd1.5, rd1.6, header = FALSE, type = "text", se = list(rd1.1$rse, rd1.2$rse, rd1.3$rse, rd1.4$rse, rd1.5$rse, rd1.6$rse))

##   
## ===================================================================================================================================================================  
## Dependent variable:   
## -----------------------------------------------------------------------------------------------------------------------------------------------  
## TotalCalories TotalMonthlyHouseholdExpenditures HouseholdMembers TotalHouseholdSavings HouseholdHeadIsEducated AdultMales\_inthehousehold  
## (1) (2) (3) (4) (5) (6)   
## -------------------------------------------------------------------------------------------------------------------------------------------------------------------  
## cash -18.110 -12.246 -0.068 -72.736 -0.0002 0.011   
## (36.598) (40.472) (0.092) (166.509) (0.030) (0.041)   
##   
## credit -19.795 8.954 -0.009 -51.921 -0.011 0.039   
## (37.994) (54.438) (0.091) (181.413) (0.031) (0.041)   
##   
## info -77.994\* -61.094 0.056 192.757 -0.035 -0.004   
## (44.546) (41.351) (0.113) (228.368) (0.034) (0.049)   
##   
## Constant 2,099.301\*\*\* 1,067.080\*\*\* 3.993\*\*\* 1,418.291\*\*\* 0.252\*\*\* 1.182\*\*\*   
## (30.419) (34.527) (0.076) (134.901) (0.025) (0.034)   
##   
## -------------------------------------------------------------------------------------------------------------------------------------------------------------------  
## Observations 1,893 1,892 1,892 997 1,892 1,893   
## R2 0.002 0.001 0.001 0.002 0.001 0.001   
## Adjusted R2 0.0004 -0.001 -0.001 -0.001 -0.001 -0.001   
## Residual Std. Error 545.696 (df = 1889) 720.919 (df = 1888) 1.333 (df = 1888) 2,008.575 (df = 993) 0.429 (df = 1888) 0.602 (df = 1889)   
## ===================================================================================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

**Answer:** the equivalent table in the paper is table 1

## 3.2 **Question: How should the coefficients in the table above be interpreted? What should we look for in this table?**

**Answer:** We need to check if the coefficients are positive or negative. All the baseline household characteristics are negative relative with cash. Except TotalMonthlyHouseholdExpenditures, other characteristics are negative with credit. Except HouseholdMembers and TotalHouseholdSavings, other characteristics are negative with info. And they are all not statistics significant, which is exactly what we need. The treatment influences the consumption only through migrant by randomization.

## 3.3 **Question: Using the round 2 data, regress migrant on the treatment arm indicators. What is the equivalent table in the paper?**

**Code:**

rd2.1 <- felm(migrant ~ cash + credit + info, selectround2)  
stargazer(rd2.1, header = FALSE, type = "text", se = list(rd2.1$rse))

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## migrant   
## -----------------------------------------------  
## cash 0.230\*\*\*   
## (0.033)   
##   
## credit 0.208\*\*\*   
## (0.034)   
##   
## info -0.0004   
## (0.039)   
##   
## Constant 0.360\*\*\*   
## (0.028)   
##   
## -----------------------------------------------  
## Observations 1,871   
## R2 0.043   
## Adjusted R2 0.041   
## Residual Std. Error 0.490 (df = 1867)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

**Answer:** the equivalent table in the paper is table 2

## 3.4 **Question: How should the coefficients in the table above be interpreted? Why is this table important?**

**Answer:** Migrant is positive related with cash and credit and statistics significant but negative related with info and not statistics significant.

## 3.5 **Question: What is the underlying migration rate in the control group and how might this change our interpretation of the results?**

**Answer:** The underlying migration rate in the control group is 0.360. By dummy variable, one additional cash predicts migrant 0.230+0.360 higher, one additional credit predicts 0.208+0.360 higher, and one additional info predicts -0.0004+0.360 higher.

## 3.6 **Question: (2 pages) Replicate the results presented in the third row of the first three columns of table 3. Present them in a table and interpret these results.**

Note 1: The authors elect to drop one household observation because the reported value of total fish consumed in the household is very high.

Note 2: To replicate the standard errors in the paper you will need to cluster your standard errors at the village level. We will discuss clustering later in the semester. Using felm you can specify the level of clustering (clustervariable) using the following command:

reg<-felm(Y~x1|fevariables|ivfirststage|clustervariable, dataname)

where you can replace fevariables and ivfirststage with 0 if you are not using fixed effects or an instrument.

**Code:**

rd2 <- subset(selectround2, total\_fish < 16359)  
  
rd2.2 <- felm(average\_exp2 ~ cash + credit + info|upazila|0|village, data = rd2)  
  
stargazer(rd2.2, header = FALSE, type = "text", se = list(rd2.2$rse))

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## average\_exp2   
## -----------------------------------------------  
## cash 96.566\*\*\*   
## (27.229)   
##   
## credit 76.743\*\*\*   
## (27.995)   
##   
## info 38.521   
## (39.657)   
##   
## -----------------------------------------------  
## Observations 1,869   
## R2 0.044   
## Adjusted R2 0.036   
## Residual Std. Error 452.094 (df = 1852)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

**Answer:** Total consumption is positive related with cash, credit and info, and it is statistics significant on cash and credit.

## 3.7 **Question: What happens to these estimates if you drop the fixed effects from the specification. Why?**

**Code:**

rd2.3 <- felm(average\_exp2 ~ cash + credit + info|0|0|village, data = rd2)  
  
stargazer(rd2.3, header = FALSE, type = "text", se = list(rd2.3$rse))

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## average\_exp2   
## -----------------------------------------------  
## cash 88.051\*\*\*   
## (27.096)   
##   
## credit 49.136\*   
## (26.468)   
##   
## info -6.745   
## (37.265)   
##   
## Constant 954.133\*\*\*   
## (19.662)   
##   
## -----------------------------------------------  
## Observations 1,869   
## R2 0.007   
## Adjusted R2 0.005   
## Residual Std. Error 459.305 (df = 1865)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

**Answer:** The coefficients are became smaller and less significant. There are time fixed effects and individual fixed effect.

## 3.8 **Question: (2 pages)Replicate the results presented in the third row of the fourth and fifth columns of table 3. What happens to the coefficient and standard errors? Is this surprising? What does this tell us?**

Hint: You will need to construct a new variable to run these estimates.

**Code:**

rd2$cc = ifelse(rd2$cash == 1 | rd2$credit == 1, 1, 0)  
  
rd2.4 <- felm(average\_exp2 ~ cc|upazila|0|village, data = rd2)  
rd2.5 <- felm(average\_exp2 ~(cc+lit+walls\_good+subsistencer1+num\_adltmalesr1+num\_childrenr1+constrainedr1+bankedr1+exp\_total\_pc\_r1+monga+avgQ13earned+dhaka\_network+dhaka\_remit)|upazila|0|village, data = rd2)  
  
stargazer(rd2.4, rd2.5, header = FALSE, type = "text", se = list(rd2.4$rse, rd2.5$rse), title = "")

##   
## ===========================================================  
## Dependent variable:   
## ---------------------------------------  
## average\_exp2   
## (1) (2)   
## -----------------------------------------------------------  
## cc 68.359\*\*\* 60.139\*\*   
## (23.991) (23.488)   
##   
## lit -9.590   
## (10.952)   
##   
## walls\_good 97.810\*\*\*   
## (22.030)   
##   
## subsistencer1 -328.904\*\*\*   
## (113.356)   
##   
## num\_adltmalesr1 -40.548\*\*\*   
## (15.052)   
##   
## num\_childrenr1 -129.078\*\*\*   
## (13.439)   
##   
## constrainedr1 -47.221   
## (42.062)   
##   
## bankedr1 51.829\*\*   
## (21.141)   
##   
## exp\_total\_pc\_r1 0.075\*\*   
## (0.038)   
##   
## monga -0.279   
## (0.544)   
##   
## avgQ13earned 53.502\*\*\*   
## (13.823)   
##   
## dhaka\_network 0.419   
## (0.393)   
##   
## dhaka\_remit 0.616   
## (0.444)   
##   
## -----------------------------------------------------------  
## Observations 1,869 1,825   
## R2 0.044 0.147   
## Adjusted R2 0.036 0.134   
## Residual Std. Error 452.023 (df = 1854) 430.506 (df = 1798)  
## ===========================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

**Answer:** Both coefficients and standard errors decreased. Because the control variable changed, column 5 indicates that the effects are generally robust to adding some controls for baseline characteristics, therefore it is not surpring.

## 3.9 **Question: Why is the header of the first five columns of table 3 “ITT”. What is meant by this and what does this tell us about how we should interpret these results?**

**Answer:** ITT: intend to treat The analysis is conducted on an intention-to-treat basis, irrespective of whether households received or completed the intended treatment. In Column 4, intent-to-treat estimates for ash and credit incentive treatments are presented together. Notably, there is an observed increase in average monthly household consumption in these incentive villages, leading to additional calories per person per day. Furthermore, Column 5 demonstrates the robustness of these effects, showing that they generally persist even after incorporating some controls for baseline characteristics.

## 3.10 **Question: We are interested in estimating how migration affects total expenditures for the households that were induced to migrate by the cash and credit treatments as follows,**

**where is dummy indicator for if a member of household i in village v in subdistrict j migrated, is a vector of control variables and are the subdistrict fixed effects. However it is not possible to identify in the data which households were induced by the treatment vs those who would have migrated either way. Furthermore, there is likely substantial selection between the households that select into migration versus those that do not. Propose a source of exogenous variation that can be used as an instrument to isolate “good” exogenous variation in migration.**

**Answer:** Average skill score of network

## 3.11 **Question: What is the first stage specification?**

**Answer:** The first stage specification in the instrumental variable (IV) approach involves estimating the relationship between the instrumental variable (IV) and the endogenous variable. In this case, the endogenous variable is and IV is Average skill score of network.

## 3.12 **Question: (2 pages) Estimate the first stage and check that you have a strong instrument for migration.**

Note: The first stage results reported in the paper appendix may differ slightly as explained in the table footnote.

**Code:**

##first stage: check if instrument variable is related with X  
  
rd2.6 <- felm(migrant ~ avgQ13earned, selectround2)  
rd2.7 <- felm(migrant ~ avgQ13earned + cash + credit + info, selectround2)  
  
#stargazer(rd2.5, header = FALSE, type = "text", se = list(rd2.3$rse))  
summary(rd2.6)

##   
## Call:  
## felm(formula = migrant ~ avgQ13earned, data = selectround2)   
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8309 -0.4909 0.2341 0.4737 0.7991   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.069125 0.060594 -1.141 0.254   
## avgQ13earned 0.090001 0.009275 9.704 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4879 on 1839 degrees of freedom  
## (因为不存在，66个观察量被删除了)  
## Multiple R-squared(full model): 0.04871 Adjusted R-squared: 0.04819   
## Multiple R-squared(proj model): 0.04871 Adjusted R-squared: 0.04819   
## F-statistic(full model):94.16 on 1 and 1839 DF, p-value: < 2.2e-16   
## F-statistic(proj model): 94.16 on 1 and 1839 DF, p-value: < 2.2e-16

summary(rd2.7)

##   
## Call:  
## felm(formula = migrant ~ avgQ13earned + cash + credit + info, data = selectround2)   
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8693 -0.4628 0.1979 0.4352 0.8579   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.149507 0.063735 -2.346 0.0191 \*   
## avgQ13earned 0.081635 0.009187 8.886 < 2e-16 \*\*\*  
## cash 0.202448 0.033277 6.084 1.43e-09 \*\*\*  
## credit 0.187817 0.034237 5.486 4.69e-08 \*\*\*  
## info 0.006495 0.039537 0.164 0.8695   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.48 on 1836 degrees of freedom  
## (因为不存在，66个观察量被删除了)  
## Multiple R-squared(full model): 0.08078 Adjusted R-squared: 0.07877   
## Multiple R-squared(proj model): 0.08078 Adjusted R-squared: 0.07877   
## F-statistic(full model):40.33 on 4 and 1836 DF, p-value: < 2.2e-16   
## F-statistic(proj model): 40.33 on 4 and 1836 DF, p-value: < 2.2e-16

**Answer:** F statistics are very large (greater than 10), therefore it’s a strong instrument.

## 3.13 **Question: (2 pages) Use your instrument to estimate the LATE (Local Average Treatment Effect), the impact of migration on total consumption for those induced to migrate by the treatment, as in columns 6 and 7 of table 3 in the paper. Interpret your results.**

Note: if you wish to replicate the paper’s coefficients exactly, you will need to use multiple instruments, one for each treatment arm.

**Code:**

LATE <- felm(average\_exp2 ~ migrant|avgQ13earned, rd2)  
  
#summary(LATE)  
stargazer(LATE, header = FALSE, type = "text", se = list(LATE$rse))

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## average\_exp2   
## -----------------------------------------------  
## migrant 104.682\*\*\*   
## (23.486)   
##   
## -----------------------------------------------  
## Observations 1,837   
## R2 0.319   
## Adjusted R2 0.095   
## Residual Std. Error 439.824 (df = 1381)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

**Answer:** Migrant is positive related with total consumption, and it’s statistics significant.

## 3.14 **Question: Why are these results different from those in columns 4 and 5 of the paper?**

**Answer:** Considered instrument variable, which can reduce the influence of bias. Also the variables need additional controls.

## 3.15 **Question: Why is this value particularly relevant for policy decisions in the context of this experiment.**

**Answer:** The LATE specifically focuses on the subpopulation that is induced to migrate due to the treatment, allowing for a more precise understanding of the treatment’s effect. Also the LATE helps identify the impact of migration for those households that were influenced by the cash and credit treatments, providing targeted insights for policy making.

## 3.16 **Question: Suppose a policy maker found these results so compelling that they decided to make this a national policy. How would general equilibrium effects potentially change the impacts of this policy if it was implemented in a very large scale way?**

**Answer:** The policy may lead to changes in property values, rental prices, and housing availability. The policy could influence national labor market dynamics, leading to changes in wages and employment levels across industries and sectors. Shifts in the supply and demand for labor may vary regionally.

## 3.17 **Question: One major concern that is often brought up in discussions about RCT’s is the problem of external validity. It is not always clear how informative the findings from a small scale research project in one context are for policy makers working on a different scale and in different contexts. What are your thoughts on the external validity of this particular project and RCT’s in general?**

**Answer:** The participants in RCTs are typically a specific sample, and the extent to which the results apply to a broader population depends on the diversity of the participants. If the study sample is not representative, external validity may be limited.

# 4. Submission instructions:

* Make sure the final version of your assignment is knit in pdf format and uploaded to gradescope. Make sure you have one question response per page (unless otherwise indicated) so that question positions align with the template in gradescope.The final PDF should be 25 pages long.