Problem Set 4: Randomized Control Trials

Claire Duquennois

Name:

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:

```
#install.packages('haven',repos = "http://cran.us.r-project.org")
#install.packages("here",repos = "http://cran.us.r-project.org")
#install.packages("dplyr",repos = "http://cran.us.r-project.org")

library(haven)
library(dplyr)
library(stargazer)
library(lfe)
library(ggplot2)

datar1<-read_dta("../../../PSdata/PS4_RCTdata/data/Round1_Controls_Table1.dta")
datar1<-as.data.frame(datar1)
nrow(datar1)
```

```
## [1] 1900
```

```
datar2<-read_dta("../../../PSdata/PS4_RCTdata/data/Round2.dta")
datar2<-as.data.frame(datar2)
nrow(datar2)
```

```
## [1] 1907
```

The data for round one contains 1900 observations and the data for round two contains 1907 observations. In both datasets, the observations are at the household level. The second dataset contains more observations because of split-off households.

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:

```
datar1<-datar1[, c("cash","credit","info","control","exp_total_pc_r1","q9pdcalq9",
                   "hhmembers_r1","tsaving_hh_r1","hhh_education","num_adltmalesr1")]

datar2<-datar2[, c("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: $cash_i$, $credit_i$ and $info_i$, and present your results in a table. What is the equivalent table in the paper?

Code:

```
balangereg1<-felm(exp_total_pc_r1~cash+credit+info, datar1)
balangereg2<-felm(q9pdcalq9~cash+credit+info, datar1)
balangereg3<-felm(hhmembers_r1~cash+credit+info, datar1)
balangereg4<-felm(tsaving_hh_r1~cash+credit+info, datar1)
balangereg5<-felm(hhh_education~cash+credit+info, datar1)
balangereg6<-felm(num_adltmalesr1~cash+credit+info, datar1)


stargazer(balangereg1,balangereg2,balangereg3,balangereg4,balangereg5,balangereg6, font.size="tiny")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 1:48:08 PM

Table 3:

	Dependent variable:					
	exp_total_pc_r1	q9pdcalq9	hhmembers_r1	tsaving_hh_r1	hhh_education	num_adltmalesr1
	(1)	(2)	(3)	(4)	(5)	(6)
cash	−12.246	−18.110	−0.068	−72.736	−0.0002	0.011
	(49.633)	(37.569)	(0.092)	(187.794)	(0.030)	(0.041)
credit	8.954	−19.795	−0.009	−51.921	−0.011	0.039
	(51.067)	(38.644)	(0.094)	(195.620)	(0.030)	(0.043)
info	−61.094	−77.994*	0.056	192.757	−0.035	−0.004
	(58.571)	(44.335)	(0.108)	(220.474)	(0.035)	(0.049)
Constant	1,067.080***	2,099.301***	3.993***	1,418.291***	0.252***	1.182***
	(41.484)	(31.401)	(0.077)	(156.367)	(0.025)	(0.035)
Observations	1,892	1,893	1,892	997	1,892	1,893
R^2	0.001	0.002	0.001	0.002	0.001	0.001
Adjusted R^2	−0.001	0.0004	−0.001	−0.001	−0.001	−0.001
Residual Std. Error	720.919 (df = 1888)	545.696 (df = 1889)	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 results presented in this table are equivalent to those presented in table 1 of the paper.

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

Answer:

The coefficient on the constant gives the mean value of that variable for the omitted group, which in this case is the control group households. The other coefficients give the mean difference between household in the other treatment arm as compared to the control group households. Looking at total household consumption for instance, we see that the mean for the control group is 1067.08 (Takkas per person per month, I believe) while the mean for the cash treatment group is 1067.08-12.25=1054.83. This information is equivalent to the information presented in table 1, just presented in a different way. As with table 1, we are interested in seeing if the randomization of households between the different control and treatment groups was done correctly. We want to make sure that the different treatment arms and the control group look similar in terms of their baseline characteristics. Since almost all the estimated coefficients are small and not statistically significant, the different treatment arms are balanced on observables. There is a statistically significant difference between the information treatment and the control group on total calories consumed. This is likely due to random chance and not a significant concern. Since we are testing 18 coefficients, the laws of probability would suggest that random chance would yield a significant difference on one or two coefficients. We can nevertheless keep this baseline difference in mind when we interpret our treatment effects.

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

Code:

```
regtakeup<-felm(migrant~cash+credit+info, datar2)


stargazer(regtakeup)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 1:48:08 PM

Table 4:

	Dependent variable:
	migrant
cash	0.230^{***}
	(0.034)
credit	0.208^{***}
	(0.035)
info	-0.0004
	(0.040)
Constant	0.360^{***}
	(0.028)
Observations	1,871
R^2	0.043
Adjusted R^2	0.041
Residual Std. Error	0.490 (df = 1867)
Note:	*p<0.1; **p<0.05; ***p<0.01

Answer: These results are equivalent to those reported in row 1 of table 2.

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

Answer: The intercept coefficient gives us the migration rate for households in the control group, which is 36%. The other coefficients then tell us the difference between the control group and that particular treatment arm. Thus we can see that there is no difference between the control group and the information treatment but that the those in the incentivized treatment arms (cash and credit) are much more likely to migrate, by 23 and 21 percentage points respectively. Knowing these values is important because it allows us to think carefully about how much the treatment is actually changing behaviors: there are many households in the treatment arms that do not take up the treatment (never takers) and there are also many households in the control group who get treated (always takers), thus it will be important for us to distinguish between the average treatment effects (ATE) versus the treatment on the treated (TOT).

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 fact that 36% of the control group migrates is significant. It suggest that this migration "technology" is not novel and is a well known economic opportunity. It is thus not surprising that the information treatment had no significant effect. This also means that the people who are moved to migrate by the incentive are marginal migrants: people for whom the benefits to migrating are probably particularly low since they would not have chosen to incur the migration cost on their own.

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.

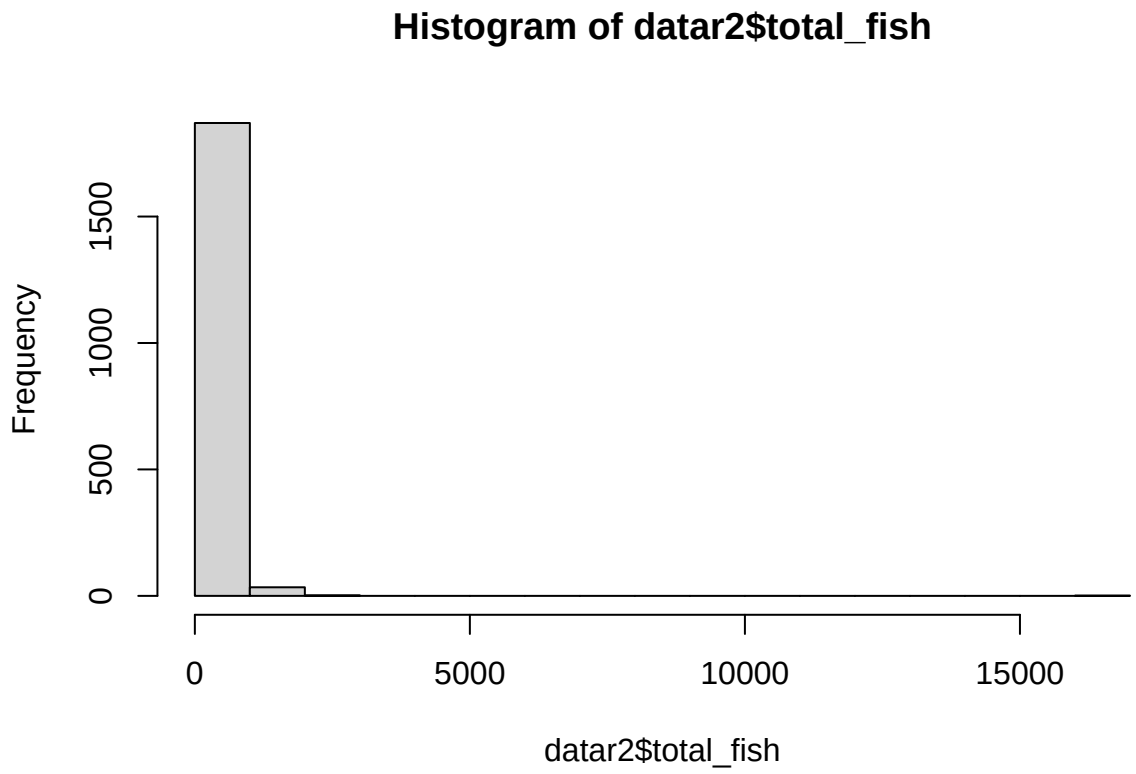
Note 2: To replicate the standard errors in the paper you will need to cluster your standard errors at the village level. We will discuss clustering later in the semester. Using `felm` you can specify the level of clustering (`clustervariable`) using the following command:

```
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:

```
#finding the outlier observation:
hist(datar2$total_fish)
```




```
summary(datar2$total_fish)
```

```
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.00    86.91   173.81   256.47   325.89 16359.82
```

```
datar2noout<-datar2[datar2$total_fish<16350, ]


regcons<-felm(average_exp2~cash+credit+info|upazila|0|village, datar2noout)
```

```
stargazer(regcons)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 1:48:08 PM

Table 5:

	Dependent variable:
	average_exp2
cash	96.566***
	(34.610)
credit	76.743**
	(33.646)
info	38.521
	(50.975)
Observations	1,869
R^2	0.044
Adjusted R^2	0.036
Residual Std. Error	452.094 (df = 1852)
Note:	*p<0.1; **p<0.05; ***p<0.01

Answer:

Among households who were offered the cash migration incentive, total consumption increased by about 97 Takka per household member compared to the control group. Those who were offered the credit treatment increased total consumption by 77 Takka. The information treatment did not have a statistically significant effect on total consumption.

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

Code:

```
regconsnofe<-felm(average_exp2~cash+credit+info|0|0|village, datar2noout)

stargazer(regconsnofe)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 1:48:08 PM

Table 6:

	Dependent variable:
	average_exp2
cash	88.051**
	(38.795)
credit	49.136
	(39.585)
info	−6.745
	(57.699)
Constant	954.133***
	(30.696)
Observations	1,869
R^2	0.007
Adjusted R^2	0.005
Residual Std. Error	459.305 (df = 1865)
Note:	*p<0.1; **p<0.05; ***p<0.01

Answer: The coefficients do change a bit but they are not statistically different from the estimates that include the fixed effects. The standard errors are also less precise. These two patterns are consistent with these results being from a randomized control trial. Because of the randomization, we do not need to be concerned with bias from omitted variables. Adding these fixed effects as controls should not change the coefficients but does allow us to get more precise estimates.

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:

```
datar2noout$incentivized<-0
datar2noout$incentivized[datar2noout$cash==1]<-1
datar2noout$incentivized[datar2noout$credit==1]<-1


regconscol4<-felm(average_exp2~incentivized|upazila|0|village, datar2noout)
regconscol5<-felm(average_exp2~incentivized+lit+walls_good+monga+dhaka_remit+dhaka_network
                  +exp_total_pc_r1+subsistencer1+num_adltmalesr1+num_childrenr1+avgQ13earned
                  +constrainedr1+bankedr1|upazila|0|village, datar2noout)

stargazer(regconscol4, regconscol5)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 1:48:08 PM

Answer: The coefficients do change a bit but they are not statistically different from the estimates without the controls and the standard errors are more precise. As above, this is consistent with these results being from a randomized control trial. Because of the randomization, we do not need to be concerned with bias from omitted variables. Adding these controls should not change the coefficients but does allow us to get more precise estimates.

Table 7:

	Dependent variable:	
	average_exp2	
	(1)	(2)
incentivized	68.359**	60.139**
	(30.593)	(29.683)
lit		−9.590
		(10.020)
walls_good		97.810***
		(24.689)
monga		−0.279
		(0.591)
dhaka_remit		0.616
		(0.397)
dhaka_network		0.419
		(0.380)
exp_total_pc_r1		0.075*
		(0.039)
subsistencer1		−328.904***
		(116.103)
num_adltmalesr1		−40.548***
		(13.851)
num_childrenr1		−129.078***
		(13.285)
avgQ13earned		53.502***
		(16.184)
constrainedr1		−47.221
		(41.392)
bankedr1		51.829**
		(22.694)
Observations	1,869	1,825
R^2	0.044	0.147
Adjusted R^2	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

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 stands for "Intent to Treat" since these are intent-to-treat estimates. These estimates tell us the average effect on all households that were randomized into that treatment arm, whether or not the household actually took up treatment and sent a migrant to the city. Based on the take up information we looked at earlier, this average reflect the outcome of several different types of households: a) never taker households (who's outcome is unchanged by treatment), b) always taker households (who's outcome is unchanged by treatment) and c) compliers (who are moved to migrate by treatment). Since, through randomization, our control group also includes these same types of households, any difference we observe is entirely driven by the difference in outcomes for the compliers, which are about 22% of the households. Thus these households must experience a very large treatment effect to generate such a large ITT treatment effect.

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,

$$TotExp_{ivj} = \alpha + \beta_1 Migrate_{ivj} + \theta X_{ivj} + \varphi_j + \nu_{ivj}$$

where $Migrate_{ivj}$ is dummy indicator for if a member of household i in village v in subdistrict j migrated, X_{ivj} is a vector of control variables and φ_j 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: We exogenous variation generated by the randomization from the RCT as an instrument for Migration.

3.11 Question: What is the first stage specification?

Answer:

$$Migrant_{ivj} = \lambda + \rho Z_{ivj} + \gamma X_{ivj} + \varphi_j + \epsilon_{ivj}$$

where Z_{ivj} is an indicator for the treatment arms and the other variables are as defined above.

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:

```
regfs1<-felm(migrant_new~cash+credit+info|upazila|0|village, datar2noout)
regfs2<-felm(migrant_new~cash+credit+info+lit+walls_good+monga+dhaka_remit+dhaka_network
                  +exp_total_pc_r1+subsistencer1+num_adltmalesr1+num_childrenr1+avgQ13earned
                  +constrainedr1+bankedr1|upazila|0|village, datar2noout)


stargazer(regfs1,regfs2)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 1:48:08 PM

```
summary(regfs1)$fstat
```

[1] 13.16932

```
summary(regfs2)$fstat
```

[1] 10.90831

Answer: The f stat for both specifications is greater than 10 which passes the benchmark for a sufficiently strong instrument. Furthermore, because we know that this variation is explicitly random, violations of the exclusion restriction are less of a concern here, unless there is reason to believe that receiving a certain treatment may have affected consumption in a way other than through the migration mechanism.

Table 8:

	Dependent variable:	
	migrant_new	
	(1)	(2)
cash	0.170***	0.181***
	(0.044)	(0.044)
credit	0.166***	0.168***
	(0.043)	(0.044)
info	−0.011	0.002
	(0.044)	(0.044)
lit		0.011
		(0.012)
walls_good		−0.032
		(0.023)
monga		0.0005
		(0.001)
dhaka_remit		0.001**
		(0.0004)
dhaka_network		0.0001
		(0.0004)
exp_total_pc_r1		−0.00001
		(0.00001)
subsistencer1		−0.518***
		(0.123)
num_adltmalesr1		0.086***
		(0.019)
num_childrenr1		0.034***
		(0.013)
avgQ13earned		0.036*
		(0.020)
constrainedr1		−0.066
		(0.059)
bankedr1		0.065***
		(0.025)
Observations	1,871	1,827
R^2	0.102	0.145
Adjusted R^2	0.094	0.132
Residual Std. Error	0.463 (df = 1854)	0.454 (df = 1798)
Note:		*p<0.1; **p<0.05; ***p<0.01

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:

```
regiv1<-felm(average_exp2~1|upazila|(migrant_new~cash+credit+info)|village, datar2noout)
regiv2<-felm(average_exp2~lit+walls_good+monga+dhaka_remit+dhaka_network
                  +exp_total_pc_r1+subsistencer1+num_adltmalesr1+num_childrenr1+avgQ13earned
                  +constrainedr1+bankedr1
             |upazila
             |(migrant_new~cash+credit+info+lit+walls_good+monga+dhaka_remit+dhaka_network
               +exp_total_pc_r1+subsistencer1+num_adltmalesr1
                  +num_childrenr1+avgQ13earned+constrainedr1
               +bankedr1)|village,datar2noout)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

```
stargazer(regiv1,regiv2)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 1:48:09 PM

Answer:

Total consumption in households that were induced to migrate by these treatments increased by about 355-391 takka depending on the specification. This is substantial since the mean of total consumption is about 1000 takka.

Table 9:

	Dependent variable:	
	average_exp2	
	(1)	(2)
lit		−13.839
		(10.403)
walls_good		109.154***
		(23.546)
monga		−0.454
		(0.568)
dhaka_remit		0.310
		(0.462)
dhaka_network		0.381
		(0.404)
exp_total_pc_r1		0.078*
		(0.040)
subsistencer1		−152.100
		(147.358)
num_adltmalesr1		−71.515***
		(18.559)
num_childrenr1		−141.310***
		(13.134)
avgQ13earned		40.273***
		(14.391)
constrainedr1		−23.708
		(43.510)
bankedr1		28.396
		(27.692)
‘migrant_new(fit)‘	391.193**	355.115**
	(170.926)	(160.785)
Observations	1,869	1,825
R^2	0.011	0.146
Adjusted R^2	0.004	0.133
Residual Std. Error	459.639 (df = 1854)	430.750 (df = 1798)
Note:		*p<0.1; **p<0.05; ***p<0.01

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

Answer: These values tell us the LATE, the estimated effect on the households who are actually affected by the treatment. The results in columns 4 and 5 tell us the average effect of being in an incentivized village, but many households do not change their migration decision, so the effect is diluted by the non-compliance of never takers.

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

Answer: In this experiment, the LATE is a very relevant value if we think about costs and benefits. For this policy, the policymaker only needs to incur the cost of the policy for individuals who choose to migrate. This will be the always takers and the compliers. The benefits of the policy will be the increased consumption by the compliers. Thus the values we just calculated will be key to determining if the policy is cost effective.

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: One critique of randomized control trials is that they usually tend to be fairly small scale experiments and thus make it difficult to estimate how the proposed policy might actually play out if it was implemented at scale. In this particular example there are many general equilibrium effects that would be of concern. If this policy was widely adopted, leading to a substantial increase in seasonal migration, we might expect to see important price adjustments that could substantially shift the returns experienced by migrants as well as the sending and receiving communities. Wages in the sending and receiving areas are of particular concern, and the prices of certain consumption goods could adjust as well. If migrant wages in the city, or the risk of unemployment in the city, changes substantially due to a large influx of additional migrants, the returns to migrating could easily be reduced to zero. In the sending villages, wages could increase due to the scarcity of labor. The authors explore there general equilibrium wage effects in a subsequent project.

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:

No "correct" answer. We would like to see your opinion on this issue.

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.