PAPE tutorial #02

In this session, we will explore the analysis of RCTs. The goal will be to replicate the main results of the paper by Dupas and Robinson (2013). To do so, you first need to download the replication package of their paper on the ICPSR repository. Once you have done so, you need to unzip the folder you downloaded and put the files it contains in the working directory you want to use.

You should now be able to open the dataset.

```
## set your working directory
setwd("your working directory")

## load library
library(haven)

## load dataset
df <- read_stata("HARP_ROSCA_final.dta")
```

1. Read the introduction of the paper to understand what it is about.
2. In the paper, the randomisation is done at the ROSCA level. Why isn't it done at the individual level according to you?

There are multiple reasons to randomize at the ROSCA level:

- Minimizes contamination (spillovers),
- Necessary because of the social nature of some treatments,
- Ethics and logistics.

3. In the paper, you can see that the randomization was stratified at the ROSCA level. What does it mean? Why is it interesting to use stratified randomization?

Stratification means that we randomize within a set of people who have some characteristics in common. In the case of this paper, @dupas2013 stratify on the following characteristics: gender composition, meeting frequency, and whether the ROSCA provided loans to its members. These are continuous variables. According to you, how do they manage to stratify?

Stratification ensures balance on the variables you stratify on. If these variables are important for the outcomes, it can increase the precision of your estimate.

4. Read the replication package documentation. It details all variables included in the dataset. Cite a 2-3 variables that pertain to the following category: treatment assignment, take-up of treatment, outcomes.

5. In the dataset, you have treatment indicators at the group (ROSCA) level, and at the individual level. Why was it important to include both of these for the authors?

- Individuals can be part of multiple ROSCAs, and thus be treated more than once.
- It is nicer to have dummies for the variables to be included in the regression, and given point (i) it is more convenient to directly use dummies at the individual level.
- More convenient since outcome is measured at the individual level.

6. Compute the intent-to-treat effect for all of the treatments on the amount spent on preventive health products since baseline. Also replicate the results of the papers adding to the regression control variables: strata dummies, monthly ROSCA contribution, indicator for receiving multiple treatments.

Beware: Chaisemartin and Ramirez-Cuellar (2024) about the inlusion of strata and estimation of ATE. Mention Goldsmith-Pinkham, Hull, and Kolesár (2024) about contamination bias. Mention Lin (2013) about interacted regression.

```
## load library for estimation with cluster robust covariance
library(estimatr)

## perform regressions
lm_nocov <- lm_robust(data = df,
          formula = fol2_amtinvest_healthproducts ~ safe_box +
            locked_box +
            health_pot +
            health_savings,
          clusters = id_harp_rosca)

lm_cov <- lm_robust(data = df,
          formula = fol2_amtinvest_healthproducts ~ safe_box +
            locked_box +
            health_pot +
            health_savings +
```

```
            multitreat +
            rosbg_monthly_contrib +
            as.factor(strata),
          clusters = id_harp_rosca)

## show regressions in a table
library(texreg)
```

```
Version:  1.39.3
Date:     2023-11-09
Author:   Philip Leifeld (University of Essex)

Consider submitting praise using the praise or praise_interactive functions.
Please cite the JSS article in your publications -- see citation("texreg").


Attachement du package : 'texreg'

L'objet suivant est masqué depuis 'package:tidyr':

    extract
```

```
# Create readable variable labels for the table
custom_names <- list(
  "safe_box" = "Safe Box",
  "locked_box" = "Locked Box",
  "health_pot" = "Health Pot",
  "health_savings" = "Health Savings account",
  "(Intercept)" = "Control Mean"
)

# Generate LaTeX table

texreg(
  list(lm_nocov,
       lm_cov),
  custom.model.names = c("No covariates",
                         "ROSCA covariates"),
  custom.coef.map = custom_names,
  omit.coef = c("as.factor\\(strata\\)", "multitreat", "rosbg_monthly_contrib"),
  stars = c(0.01, 0.05, 0.1),
```

```
  float.pos = "H",
  caption = "ITTs",
  include.ci = FALSE
)
```

	No covariates	ROSCA covariates
Safe Box	55.82	193.85**
	(57.89)	(87.58)
Locked Box	−52.28	64.84
	(51.96)	(72.76)
Health Pot	230.02**	356.33***
	(86.33)	(110.12)
Health Savings account	−53.98	33.70
	(49.73)	(66.63)
Control Mean	341.00***	409.63**
	(39.74)	(182.39)
R^2	0.03	0.06
Adj. R^2	0.03	0.03
Num. obs.	771	771
RMSE	578.09	577.22
N Clusters	112	112

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 1: ITTs

7. Interpret the coefficients and discuss what we learn from the ITT.

The ITTs effects are informative about what happens when we propose the treatment (vs when people receive it). In our case, the ITTs show that proposing social pressure towards preventive investments works very well, with nearly a doubling of the investments in preventive health prodcuts induced by the proposition to receive the treatment. Proposing a safe box also seems efficient although less than the health pot. By contrast, the lockbox and the HSA treatment do not seem effective (note that HSA was not designed for preventive health investments). However, to make more sense of these results, we also need to check the take-up of the different treatments.

8. Compute take-up for each of the treatments measured at the second follow-up, and explain why we should care about it.

Mention Chaisemartin and Ramirez-Cuellar (2024) about the inlusion of strata and estimation of ATE. Mention Goldsmith-Pinkham, Hull, and Kolesár (2024) about contamination bias. Mention Lin (2013) about interacted regression.

```r
follow_up_vars <- c("fol1_uses_safe_box",
                    "fol1_uses_lbox",
                    "fol1_hp_pot_contribute",
                    "fol1_hsa_deposit_hsa")


# Compute means for the specified variables
means <- colMeans(df[follow_up_vars], na.rm = TRUE)

# Create LaTeX table with the computed means
table <- sprintf("
\\begin{table}[ht]
\\centering
\\begin{tabular}{l|c|c|c|c}
\\hline
 & Safe Box & Lock Box & Health Pot & Health Account Savings \\\\
\\hline
Take-Up & %.2f & %.2f & %.2f & %.2f \\\\
\\hline
\\end{tabular}
\\caption{Means of Take-Up Variables}
\\label{tab:take_up_means}
\\end{table}
",
  means['fol1_uses_safe_box'],
  means['fol1_uses_lbox'],
  means['fol1_hp_pot_contribute'],
  means['fol1_hsa_deposit_hsa']
)

# Print the LaTeX table
cat(table)
```

	Safe Box	Lock Box	Health Pot	Health Account Savings
Take-Up	0.74	0.65	0.65	0.93

Table 2: Means of Take-Up Variables

9. In the dataset, you will also see that there is attrition in the follow-up data. What is attrition? Why does attrition matter?

Attrition refers to the loss of units in the sample between the treatment assignment and the

outcome measure, meaning units who (i) got assigned to the treatment and (ii) who were not observed at endline. Attrition matters because:

- It decreases statistical precision (power);
- It can threaten both the internal (differential attrition) and external (similar in the two groups but non random attrition) validity of the study.

10. Compute the attrition rate in the first follow-up survey in all treatments groups.

```
lm_f1 <- lm(data = df,
            formula = has_followup1 ~ safe_box +
              locked_box +
              health_pot +
              health_savings+
              multitreat)

lm_f2 <- lm(data = df,
            formula = has_followup2 ~ safe_box +
              locked_box +
              health_pot +
              health_savings+
              multitreat)


## show regressions in a table
library(texreg)

# Create readable variable labels for the table
custom_names <- list(
  "safe_box" = "Safe Box",
  "locked_box" = "Locked Box",
  "health_pot" = "Health Pot",
  "health_savings" = "Health Savings account",
  "(Intercept)" = "Control Mean"
)

# Generate LaTeX table

texreg(
  list(lm_f1,
       lm_f2),
  custom.model.names = c("Follow-up 1",
                         "Follow-up 2"),
```

```
  omit.coef = c("multitreat"),
  custom.coef.map = custom_names,
  stars = c(0.01, 0.05, 0.1),
  float.pos = "h",
  caption = "Attrition in the experimental groups",
  include.ci = FALSE
)
```

	Follow-up 1	Follow-up 2
Safe Box	−0.05	−0.00
	(0.04)	(0.03)
Locked Box	−0.01	0.02
	(0.04)	(0.03)
Health Pot	−0.03	−0.03
	(0.04)	(0.03)
Health Savings account	−0.07**	0.02
	(0.04)	(0.03)
Control Mean	0.94***	0.92***
	(0.03)	(0.02)
R^2	0.24	0.01
Adj. R^2	0.24	0.00
Num. obs.	833	833

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$

Table 3: Attrition in the experimental groups

Note that a full analysis would also test for differential attrition across treatment arms, and based on individual and ROSCA characteristics.

11. Compute the local average treatment effect of each of the treatments on the amount spent on preventive health products since baseline. Use the Wald ratio.

To use the Wald ratio, we just need the quantities we previously computed. You will find an example below for the safe box arm:

```
itt <- mean(df[df$safe_box==1 & df$locked_box==0 & df$health_pot==0 & df$health_savings==0,]$
  mean(df[df$safe_box==0 & df$locked_box==0 & df$health_pot==0 & df$health_savings==0,]$fol2_

take_up <- mean(df[df$safe_box==1 & df$locked_box==0 & df$health_pot==0 & df$health_savings==

late <- itt/ take_up
```

We see that the LATE is 205.0658454.
In my coding, the LATE is in some sense "strongly" local. Why? Moreover, in this code, I do not remove from the take-up the average number of individuals in the control group who use the safe box. Why?

12. Repeat the previous question, but use an IV strategy. What are the relative benefits of the Wald estimator and of the IV regression?

```
## correct variables to fit in the model
df$fol2_uses_safe_box <- ifelse(is.na(df$fol2_uses_safe_box),
                                0,
                                1)

df$fol2_uses_lbox <- ifelse(is.na(df$fol2_uses_lbox),
                                0,
                                1)

df$fol2_hp_pot_contribute <- ifelse(is.na(df$fol2_hp_pot_contribute),
                                0,
                                1)

df$fol2_hsa_deposit_hsa <- ifelse(is.na(df$fol2_hsa_deposit_hsa),
                                0,
                                1)

df$strata <- as.factor(df$strata)

## run model

IV_cov <- iv_robust(data = df,
          formula = fol2_amtinvest_healthproducts ~ safe_box +
            locked_box +
            health_pot +
            health_savings +
            multitreat +
            rosbg_monthly_contrib +
            strata | fol2_uses_safe_box +
                                fol2_uses_lbox +
                                fol2_hp_pot_contribute +
                                fol2_hsa_deposit_hsa +
                                multitreat +
                                rosbg_monthly_contrib +
                                strata,
```

```
          clusters = id_harp_rosca)



# Create readable variable labels for the table
custom_names <- list(
  "safe_box" = "Safe Box",
  "locked_box" = "Locked Box",
  "health_pot" = "Health Pot",
  "health_savings" = "Health Savings account",
  "(Intercept)" = "Control Mean"
)

# Generate LaTeX table

texreg(
  list(IV_cov),
  custom.model.names = c("IV regression"),
  custom.coef.map = custom_names,
  omit.coef = c("multitreat", "rosbg_monthly_contrib"),
  stars = c(0.01, 0.05, 0.1),
  float.pos = "H",
  caption = "LATE",
  include.ci = FALSE
)
```

	IV regression
Safe Box	194.34
	(119.07)
Locked Box	110.88
	(130.17)
Health Pot	434.34***
	(155.19)
Health Savings account	46.34
	(134.45)
Control Mean	372.96*
	(189.18)
R^2	0.06
Adj. R^2	0.03
Num. obs.	771
RMSE	577.95
N Clusters	112

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 4: LATE

13. Dupas and Robinson (2013) do not compute the LATE. Why according to you?

- Relatively high take-up;
- Potential violation of exclusion restriction;
- ITT is really the parameter of interest.

14. Extra question: can you detect any interesting differences in people who take-up vs those who don't?

References

Chaisemartin, Clément de, and Jaime Ramirez-Cuellar. 2024. "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?" *American Economic Journal: Applied Economics* 16 (1): 193–212. https://doi.org/10.1257/app.20210252.

Dupas, Pascaline, and Jonathan Robinson. 2013. "Why Don't the Poor Save More? Evidence from Health Savings Experiments." *American Economic Review* 103 (4): 1138–71. https://doi.org/10.1257/aer.103.4.1138.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2024. "Contamination Bias in Linear Regressions." *American Economic Review* 114 (12): 4015–51. https://doi.org/10.1257/aer.20221116.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics* 7 (1): 295–318. https://doi.org/10.1214/12-AOAS583.