Causal Inference

Problem Set 1 - Solution

Simon Heß

August 2023

Neighborhood-level data

Neighborhood-level data

i)

Load and familiarize yourself with the `neighborhood.dta`. Answer the following questions:

a) How many neighborhoods (observations) are there in the data set? How many are in the treatment group and how many in the control group?
b) In terms of *Number of households (baseline)*, how large are the smallest and biggest neighborhood?
c) Create a variable that measures the number of businesses per household in each neighborhood. Briefly summarize descriptive statistics for this variables in a suitable plot.

Loading the data

```
library(haven)
library(dplyr)
neighborhood <- read_dta("data/neighborhood.dta")
household_endline1 <- read_dta("data/household_endline1.dta")
household_endline2 <- read_dta("data/household_endline2.dta")
```

Loading the data

```
library(haven)
library(dplyr)
neighborhood <- read_dta("data/neighborhood.dta")
household_endline1 <- read_dta("data/household_endline1.dta")
household_endline2 <- read_dta("data/household_endline2.dta")
```

i) a) number of neighborhoods

total

```
n_distinct(neighborhood$areaid)
```

i) a) number of neighborhoods

total

```
n_distinct(neighborhood$areaid)
```

```
## [1] 104
```

i) a) number of neighborhoods

total

```
n_distinct(neighborhood$areaid)
```

```
## [1] 104
```

in the treatment group

```
nrow(neighborhood[neighborhood$treatment==1,])
```

i) a) number of neighborhoods

total

```
n_distinct(neighborhood$areaid)
```

```
## [1] 104
```

in the treatment group

```
nrow(neighborhood[neighborhood$treatment==1,])
```

```
## [1] 52
```

i) a) number of neighborhoods

total

```
n_distinct(neighborhood$areaid)
```

```
## [1] 104
```

in the treatment group

```
nrow(neighborhood[neighborhood$treatment==1,])
```

```
## [1] 52
```

in the control group

```
nrow(neighborhood[neighborhood$treatment==0,])
```

i) a) number of neighborhoods

total

```
n_distinct(neighborhood$areaid)
```

```
## [1] 104
```

in the treatment group

```
nrow(neighborhood[neighborhood$treatment==1,])
```

```
## [1] 52
```

in the control group

```
nrow(neighborhood[neighborhood$treatment==0,])
```

```
## [1] 52
```

b) smallest and largest neighborhoods

```
library(modelsummary)
datasummary(area_pop_base ~ N + min + mean + max, data=neighborhood)
```

b) smallest and largest neighborhoods

```
library(modelsummary)
datasummary(area_pop_base ~ N + min + mean + max, data=neighborhood)
```

	N	min	mean	max
area_pop_base	104	46.00	262.92	555.00

b) smallest and largest neighborhoods

```
library(modelsummary)
datasummary(area_pop_base ~ N + min + mean + max, data=neighborhood)
```

	N	min	mean	max
area_pop_base	104	46.00	262.92	555.00

alternative

```
neighborhood$treated <- ifelse(neighborhood$treatment, "treat", "control")
datasummary(area_pop_base ~ treated * N + min + mean + max, data=neighborhood)
```

b) smallest and largest neighborhoods

```
library(modelsummary)
datasummary(area_pop_base ~ N + min + mean + max, data=neighborhood)
```

	N	min	mean	max
area_pop_base	104	46.00	262.92	555.00

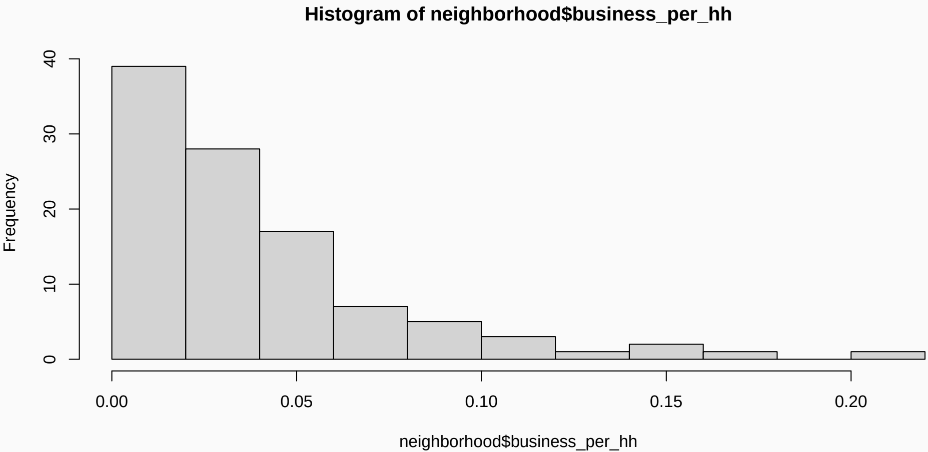
alternative

```
neighborhood$treated <- ifelse(neighborhood$treatment, "treat", "control")
datasummary(area_pop_base ~ treated * N + min + mean + max, data=neighborhood)
```

	control	treat			
	N	N	min	mean	max
area_pop_base	52	52	46.00	262.92	555.00

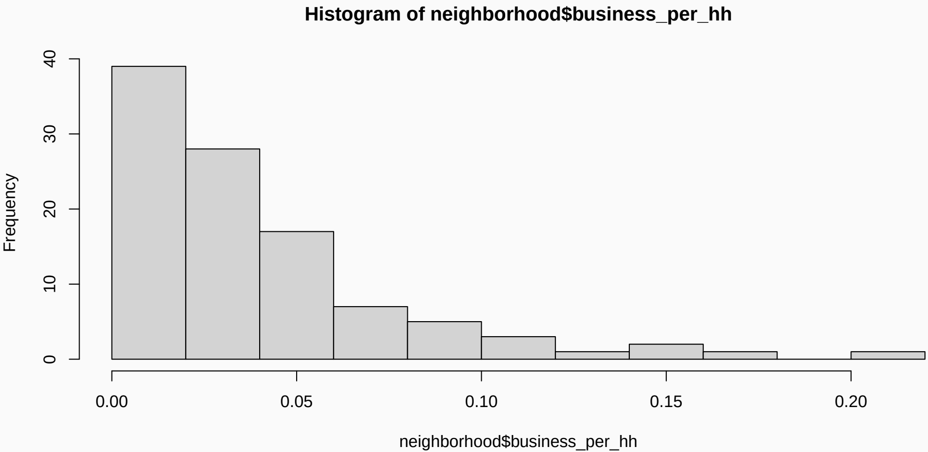
c) business per household

```
neighborhood <- neighborhood |>
  mutate(business_per_hh = area_business_total_base / area_pop_base)
hist(neighborhood$business_per_hh)
```



c) business per household

```
neighborhood <- neighborhood |>
  mutate(business_per_hh = area_business_total_base / area_pop_base)
hist(neighborhood$business_per_hh)
```



ii) balance tests

ii)

Create a table showing the means of all variables named `area_*` and the variable that you generated in **i.c)** for two sub-samples: the control group and the treatment group. Add a column to the table showing the results of individual t-tests for whether each of the variables differs between the control and the treatment group. Report the p-values for all the tests. Give a concise interpretation of the results in the table. What can we learn from them?

ii) balance tests

```
datasummary_balance( ~ treated,
  data = neighborhood |>
    subset(
      select=c("area_pop_base", "area_business_total_base", "area_debt_total_base",
               "area_exp_pc_mean_base", "area_literate_head_base", "area_literate_base",
               "business_per_hh", "treated")),
  dinm_statistic="p.value")
```

ii) balance tests

```
datasummary_balance( ~ treated,
  data = neighborhood |>
    subset(
      select=c("area_pop_base", "area_business_total_base", "area_debt_total_base",
               "area_exp_pc_mean_base", "area_literate_head_base", "area_literate_base",
               "business_per_hh", "treated")),
  dinm_statistic="p.value")
```

	control (N=52)		treat (N=52)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p
area_pop_base	264.6	160.5	261.2	142.8	-3.4	0.910
area_business_total_base	7.3	5.0	6.9	5.0	-0.3	0.726
area_debt_total_base	39675.3	47776.8	32694.1	17755.5	-6981.2	0.327
area_exp_pc_mean_base	1005.0	171.5	1047.8	195.7	42.8	0.238
area_literate_head_base	0.6	0.2	0.6	0.1	0.0	0.811
area_literate_base	0.7	0.1	0.7	0.1	0.0	0.976
business_per_hh	0.0	0.0	0.0	0.0	0.0	0.704

ii) balance tests

- neither of the baseline characteristics differs significantly between treatment and control areas

ii) balance tests

- neither of the baseline characteristics differs significantly between treatment and control areas
- this is important

ii) balance tests

- neither of the baseline characteristics differs significantly between treatment and control areas
- this is important
  - otherwise differences in outcomes might be due to differences in these confounders

ii) balance tests

- neither of the baseline characteristics differs significantly between treatment and control areas
- this is important
  - otherwise differences in outcomes might be due to differences in these confounders
- this is not surprising

ii) balance tests

- neither of the baseline characteristics differs significantly between treatment and control areas
- this is important
  - otherwise differences in outcomes might be due to differences in these confounders
- this is not surprising
  - since treatment was randomized

ii) balance tests

- neither of the baseline characteristics differs significantly between treatment and control areas
- this is important
  - otherwise differences in outcomes might be due to differences in these confounders
- this is not surprising
  - since treatment was randomized
- if this showed a large number of unbalanced characteristics
  - we would be worried about mistakes in the experiment or data collection

merging

Merge the data

Merge the `neighborhood` dataset with `household_endline1` and `household_endline2`. This merged dataset will be used to analyze the treatment effects described in Banerjee et al. (2015).

merging

Merge the data

Merge the `neighborhood` dataset with `household_endline1` and `household_endline2`. This merged dataset will be used to analyze the treatment effects described in Banerjee et al. (2015).

```
full_data_endline1 <- merge(neighborhood, household_endline1, by = "areaid")
full_data <- merge(full_data_endline1, household_endline2)
```

merging

Merge the data

Merge the `neighborhood` dataset with `household_endline1` and `household_endline2`. This merged dataset will be used to analyze the treatment effects described in Banerjee et al. (2015).

```
full_data_endline1 <- merge(neighborhood, household_endline1, by = "areaid")
full_data <- merge(full_data_endline1, household_endline2)
```

Treatment effect: Access to microcredit

iii)

Run 4 OLS regression, using the variables spandana_1, anyloan_1, spandana_2, and anyloan_2 as dependent variables. Use treatment and these six area-level control variables as independent variables in all regressions: area_pop_base, area_literate_base, area_debt_total_base, area_business_total_base, area_exp_pc_mean_base, area_literate_head_base. Cluster your standard errors at the area level and weight the regressions to account for oversampling of Spandana borrowers, i.e., use the weights w1 for endline 1 and w2 for endline 2.

a) Show your 4 estimation results in a single table. Restrict your table to show only output that is relevant to discuss the effects of microfinance. Describe and interpret your results. What is the effect of access to microcredit in treated areas? Compare the estimated effect size against the mean of the dependent variable in the control group. (<150 words)

```
models <- list(
  "Spandana 1" = lm(spandana_1 ~ treatment + area_pop_base + area_literate_base + area_debt_t
              weights = w1, data = full_data),
  "Anyloan 1" = lm(anyloan_1 ~ treatment + area_pop_base + area_literate_base + area_debt_tot
              weights = w1, data = full_data),
  "Spandana 2" = lm(spandana_2 ~ treatment + area_pop_base + area_literate_base + area_debt_t
              weights = w2, data = full_data),
  "Anyloan 2" = lm(anyloan_2 ~ treatment + area_pop_base + area_literate_base + area_debt_tot
              weights = w2, data = full_data) )
modelsummary(models, vcov = ~areaid, coef_omit = "^(?!.*treat)", gof_omit = "^(?!.*DV|.*Obs)"
```

```r
models <- list(
  "Spandana 1" = lm(spandana_1 ~ treatment + area_pop_base + area_literate_base + area_debt_t
            weights = w1, data = full_data),
  "Anyloan 1" = lm(anyloan_1 ~ treatment + area_pop_base + area_literate_base + area_debt_tot
            weights = w1, data = full_data),
  "Spandana 2" = lm(spandana_2 ~ treatment + area_pop_base + area_literate_base + area_debt_t
            weights = w2, data = full_data),
  "Anyloan 2" = lm(anyloan_2 ~ treatment + area_pop_base + area_literate_base + area_debt_tot
            weights = w2, data = full_data) )
modelsummary(models, vcov = ~areaid, coef_omit = "^(?!.*treat)", gof_omit = "^(?!.*DV|.*Obs)"
```

	Spandana 1	Anyloan 1	Spandana 2	Anyloan 2
treatment	0.127	−0.022	0.063	0.000
	(0.020)	(0.014)	(0.018)	(0.010)
Num.Obs.	6811	6862	6142	6142
Control mean DV	0.051	0.867	0.111	0.904

	Spandana 1	Anyloan 1	Spandana 2	Anyloan 2
treatment	0.127	−0.022	0.063	0.000
	(0.020)	(0.014)	(0.018)	(0.010)
Num.Obs.	6811	6862	6142	6142
Control mean DV	0.051	0.867	0.111	0.904

	Spandana 1	Anyloan 1	Spandana 2	Anyloan 2
treatment	0.127	−0.022	0.063	0.000
	(0.020)	(0.014)	(0.018)	(0.010)
Num.Obs.	6811	6862	6142	6142
Control mean DV	0.051	0.867	0.111	0.904

- At endline 1, the probability to have a Spandana loan increased due treatment by 12.7 p.p., (from 5%)

	Spandana 1	Anyloan 1	Spandana 2	Anyloan 2
treatment	0.127	−0.022	0.063	0.000
	(0.020)	(0.014)	(0.018)	(0.010)
Num.Obs.	6811	6862	6142	6142
Control mean DV	0.051	0.867	0.111	0.904

- At endline 1, the probability to have a Spandana loan increased due treatment by 12.7 p.p., (from 5%)
- Probability to have any loan did not change significantly

	Spandana 1	Anyloan 1	Spandana 2	Anyloan 2
treatment	0.127	−0.022	0.063	0.000
	(0.020)	(0.014)	(0.018)	(0.010)
Num.Obs.	6811	6862	6142	6142
Control mean DV	0.051	0.867	0.111	0.904

- At endline 1, the probability to have a Spandana loan increased due treatment by 12.7 p.p., (from 5%)
- Probability to have any loan did not change significantly
  - suggests people did not get more indebted

	Spandana 1	Anyloan 1	Spandana 2	Anyloan 2
treatment	0.127	−0.022	0.063	0.000
	(0.020)	(0.014)	(0.018)	(0.010)
Num.Obs.	6811	6862	6142	6142
Control mean DV	0.051	0.867	0.111	0.904

- At endline 1, the probability to have a Spandana loan increased due treatment by 12.7 p.p., (from 5%)
- Probability to have any loan did not change significantly
  - suggests people did not get more indebted
- By endline 2, these differences have are smaller

	Spandana 1	Anyloan 1	Spandana 2	Anyloan 2
treatment	0.127	−0.022	0.063	0.000
	(0.020)	(0.014)	(0.018)	(0.010)
Num.Obs.	6811	6862	6142	6142
Control mean DV	0.051	0.867	0.111	0.904

- At endline 1, the probability to have a Spandana loan increased due treatment by 12.7 p.p., (from 5%)
- Probability to have any loan did not change significantly
  - suggests people did not get more indebted
- By endline 2, these differences have are smaller
  - because the control communities have caught up

	Spandana 1	Anyloan 1	Spandana 2	Anyloan 2
treatment	0.127	−0.022	0.063	0.000
	(0.020)	(0.014)	(0.018)	(0.010)
Num.Obs.	6811	6862	6142	6142
Control mean DV	0.051	0.867	0.111	0.904

- At endline 1, the probability to have a Spandana loan increased due treatment by 12.7 p.p., (from 5%)
- Probability to have any loan did not change significantly
  - suggests people did not get more indebted
- By endline 2, these differences have are smaller
  - because the control communities have caught up
- Do not interpret the coefficients of control variables (especially not as causal effects)

Treatment effect: Consumption

Treatment effect: Consumption

iv)

Run 6 OLS regressions using the variables `total_exp_mo_pc_1`, `durables_exp_mo_pc_1`, `temptation_exp_mo_pc_1`, `total_exp_mo_pc_2`, `durables_exp_mo_pc_2`, `temptation_exp_mo_pc_2`, as dependent variables. As before, use the treatment dummy and the area-level controls as right-hand-side variables, cluster your standard errors at the area level and weight your regressions to account for the oversampling of Spandana borrowers.

a) Show your 6 estimation results in a single table. Restrict your table to show only output that is relevant to discuss the effects of microfinance. Describe and interpret your results. What is the effect of access to microcredit in treated areas? (<150 words) with controls:

	total 1	durables 1	temptation 1	total 2	durables 2	temptation 2
treatment	10.243	19.734	−8.785	−48.826	0.419	−10.074
	(37.217)	(11.353)	(4.915)	(51.535)	(9.876)	(6.610)
Num.Obs.	6827	6781	6827	6142	6140	6142
Control mean DV	1419.229	116.174	84.293	1914.282	155.497	117.699

- it does not seem that microfinance has significantly altered consumption patterns
- in particular, it did not increase the consumption of temptation goods

iv)

b) Reproduce the same table as in part **a)**, this time running the regressions without the area-level control variables. Do the results change qualitatively? Interpret your observations. (< 100 words)

Treatment effect: Consumption without controls

with controls:

	total 1	durables 1	temptation 1	total 2	durables 2	temptation 2
treatment	10.243	19.734	−8.785	−48.826	0.419	−10.074
	(37.217)	(11.353)	(4.915)	(51.535)	(9.876)	(6.610)
Control mean DV	1419.229	116.174	84.293	1914.282	155.497	117.699

without controls:

	total 1	durables 1	temptation 1	total 2	durables 2	temptation 2
treatment	37.980	22.432	−8.968	−30.523	1.545	−10.594
	(46.215)	(11.727)	(5.204)	(52.415)	(9.452)	(7.163)
Control mean DV	1419.229	116.174	84.293	1914.282	155.497	117.699

Treatment effect: Consumption without controls

with controls:

	total 1	durables 1	temptation 1	total 2	durables 2	temptation 2
treatment	10.243	19.734	−8.785	−48.826	0.419	−10.074
	(37.217)	(11.353)	(4.915)	(51.535)	(9.876)	(6.610)
Control mean DV	1419.229	116.174	84.293	1914.282	155.497	117.699

without controls:

	total 1	durables 1	temptation 1	total 2	durables 2	temptation 2
treatment	37.980	22.432	−8.968	−30.523	1.545	−10.594
	(46.215)	(11.727)	(5.204)	(52.415)	(9.452)	(7.163)
Control mean DV	1419.229	116.174	84.293	1914.282	155.497	117.699

- differences between the specifications are relatively small (measured by the standard errors/control means)

v)

v)

Assume someone suggests that the missing observations in `temptation_exp_mo_pc_2` may be due to selective attrition and the results are only insignificant for this reason. To investigate this, estimate treatment effect bounds that account for attrition. A very simple approach would be to:

- Create a "worst case" outcome variable that has the same values as `temptation_exp_mo_pc_2` for all non-missing values. Then impute a very high value for missing outcomes in the treatment group—say the 75th percentile of the distribution of the variable in the sample—and impute a very low value for missing outcomes in the control group—say the 25th percentile of the distribution of the variable in the sample.
- Create a "best case'' outcome variable that does the same in revers (low values for missings in the treatment group and high values for the missings in the control group.)
- Run the treatment effect regression with these two outcome variables.

```
p75 <- quantile(full_data$temptation_exp_mo_pc_2,probs = 0.75, na.rm=TRUE)
p25 <- quantile(full_data$temptation_exp_mo_pc_2,probs = 0.25, na.rm=TRUE)

full_data$temptation2_wc <- full_data$temptation_exp_mo_pc_2
full_data$temptation2_wc[full_data$treatment==1 & is.na(full_data$temptation2_wc)] <- p75
full_data$temptation2_wc[full_data$treatment==0 & is.na(full_data$temptation2_wc)] <- p25

full_data$temptation2_bc <- full_data$temptation_exp_mo_pc_2
full_data$temptation2_bc[full_data$treatment==1 & is.na(full_data$temptation2_bc)] <- p25
full_data$temptation2_bc[full_data$treatment==0 & is.na(full_data$temptation2_bc)] <- p75
```

```
p75 <- quantile(full_data$temptation_exp_mo_pc_2,probs = 0.75, na.rm=TRUE)
p25 <- quantile(full_data$temptation_exp_mo_pc_2,probs = 0.25, na.rm=TRUE)

full_data$temptation2_wc <- full_data$temptation_exp_mo_pc_2
full_data$temptation2_wc[full_data$treatment==1 & is.na(full_data$temptation2_wc)] <- p75
full_data$temptation2_wc[full_data$treatment==0 & is.na(full_data$temptation2_wc)] <- p25

full_data$temptation2_bc <- full_data$temptation_exp_mo_pc_2
full_data$temptation2_bc[full_data$treatment==1 & is.na(full_data$temptation2_bc)] <- p25
full_data$temptation2_bc[full_data$treatment==0 & is.na(full_data$temptation2_bc)] <- p75
```

	temptation 2 (worst case)	temptation 2 (best case)
treatment	4.360	-23.031
	(5.994)	(6.102)
Control mean DV	107.276	119.904

- “worst case” and “best case” scenario in terms of selective attrition
  - worst case captures what the TE could be if attrition favored smaller estimates

```
p75 <- quantile(full_data$temptation_exp_mo_pc_2,probs = 0.75, na.rm=TRUE)
p25 <- quantile(full_data$temptation_exp_mo_pc_2,probs = 0.25, na.rm=TRUE)

full_data$temptation2_wc <- full_data$temptation_exp_mo_pc_2
full_data$temptation2_wc[full_data$treatment==1 & is.na(full_data$temptation2_wc)] <- p75
full_data$temptation2_wc[full_data$treatment==0 & is.na(full_data$temptation2_wc)] <- p25

full_data$temptation2_bc <- full_data$temptation_exp_mo_pc_2
full_data$temptation2_bc[full_data$treatment==1 & is.na(full_data$temptation2_bc)] <- p25
full_data$temptation2_bc[full_data$treatment==0 & is.na(full_data$temptation2_bc)] <- p75
```

	temptation 2 (worst case)	temptation 2 (best case)
treatment	4.360	−23.031
	(5.994)	(6.102)
Control mean DV	107.276	119.904

- "worst case" and "best case" scenario in terms of selective attrition
  - worst case captures what the TE could be if attrition favored smaller estimates
  - ... so that the true effect was larger than what we find:
    - Here: 4.3 instead of -10

Bounds

- Findings:
  - even in the "worst case", the increase in temptation goods is small and insignificant
  - contrarily, in the "best case" scenario, consumption of temptation goods decreased

Bounds

- Findings:
  - even in the "worst case", the increase in temptation goods is small and insignificant
  - contrarily, in the "best case" scenario, consumption of temptation goods decreased
- Many alternative ways to estimate bounds exist

Bounds

- Findings:
  - even in the “worst case”, the increase in temptation goods is small and insignificant
  - contrarily, in the “best case” scenario, consumption of temptation goods decreased
- Many alternative ways to estimate bounds exist
  - results depend on different “worst-case” “best-case” assumptions of how attrition depends on outcomes / treatment effects

Bounds

- Findings:
  - even in the "worst case", the increase in temptation goods is small and insignificant
  - contrarily, in the "best case" scenario, consumption of temptation goods decreased
- Many alternative ways to estimate bounds exist
  - results depend on different "worst-case" "best-case" assumptions of how attrition depends on outcomes / treatment effects
  - interpretation should always link the implied assumptions to findings, so that readers can decide how credible these are

Bonus questions

vi)

Spillovers

What kind of spillovers can you imagine in the context of an RCT related to microfinance? Either think of the situation as in the paper, i.e., treatment happening at the neighborhood level, or imagine an alternative treatment in which access to microfinance is randomly allocated at the household level. Explain how the spillover in your example effects the outcome in the treatment and/or control group and how this affects the conclusions drawn from the RCT.

vi)

Spillovers

What kind of spillovers can you imagine in the context of an RCT related to microfinance? Either think of the situation as in the paper, i.e., treatment happening at the neighborhood level, or imagine an alternative treatment in which access to microfinance is randomly allocated at the household level. Explain how the spillover in your example effects the outcome in the treatment and/or control group and how this affects the conclusions drawn from the RCT.

Example 1 (treatment sharing)

- Recipients sharing with non-recipients
  - C-group has more resources, T-group less; differences smaller
  - TE estimate smaller, program appears **less** effective that it actually is

vi)

Spillovers

What kind of spillovers can you imagine in the context of an RCT related to microfinance? Either think of the situation as in the paper, i.e., treatment happening at the neighborhood level, or imagine an alternative treatment in which access to microfinance is randomly allocated at the household level. Explain how the spillover in your example effects the outcome in the treatment and/or control group and how this affects the conclusions drawn from the RCT.

Example 1 (treatment sharing)

- Recipients sharing with non-recipients
  - C-group has more resources, T-group less; differences smaller
  - TE estimate smaller, program appears **less** effective that it actually is

Example 2 (general equilibrium)

- Recipients develop businesses and attract customers
  - C-group have less customers; perform worse
  - difference between T&C larger, program appears **more** effective that it is

Appendix