Final_12265092

12265092

29/05/2022

Question 1

IA

Describe the research question in this paper in words. Explain, in words and math, the ideal experiment one might want to use to answer this question. Explain, in words and math, the naïve estimator. Provide three concrete examples of why the naïve estimator is unlikely to provide an unbiased parameter estimate in this setting, and explain whether each would bias your estimate upwards or downwards relative to the truth.

Describe the research question in this paper in words.

The main research question in the paper is whether the US Clean Water Act of 1972 effective in contributing to the reduction of water pollution.

The paper also discusses whether the benefits of the Clean Water Act exceeds its costs.

Explain, in words and math, the ideal experiment one might want to use to answer this question. Explain, in words and math, the naïve estimator.

For the US, we need to observe the following

1) Water pollution levels down stream for the plant i in US, when the plant was given grants under the Clean Water act
2) Water pollution levels down stream for the plant i in US, when the plant was not given grants under the Clean Water act

Difference between 1 and 2 gives us the effect of "Clean water Act" on the water pollution levels down stream for the plant i

Potential Outcomes framework: Let i be the individual waste water treatment plant where $i \in \{1, 2, ...N\}$. Treatment indicator D_i where $D_i \in \{0, 1\}$ Treated: $D_i = 1$: Waster water plant i was given grants under the Clean water act
Untreated: $D_i = 0$: Waster water plant i was not given grants under the Clean water act Outcome treated: $Y_i(D_i = 1)$: Water pollution downstream for plant i, when plant was given grants - Treatment Outcome untreated: $Y_i(D_i = 0)$: Water pollution downstream for plant i, when plant was not given grants - Control

We get the impact of treatment(i.e receiving grants under the Clean water act) τ_i from the difference between the above outcomes. The difference between water pollution of plant i when the plant was given grants and water pollution of plant i when the plant was not given grants

$\tau_i = Y_i(D_i = 1) - Y_i(D_i = 0)$

The impact of treatment τ_i is the difference between the two outcomes, the difference between water pollution downstreat for plant i when grants were given under the Clean Water act and the water pollution for plant i when grants were not given under the Clean Water act

From above: $\tau_i = Y_i(D_i = 1) - Y_i(D_i = 0)$

While we need both the outcomes at a given time to compute the impact of treatment, the problem is that at a given time,we can only observe either \$Y_i(D_i = 1) \$ or $Y_i(D_i = 0)$. If the plant i is treated then we observe only $Y_i(D_i = 1)$ and if the plant i is not treated then we observe only $Y_i(D_i = 0)$.

Average Treatment Effect \$ tauˆ{ATE}\$

$tau^ATE = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$

ATE measures the average effect of treatment across a population of provinces. ATE measures the effect of grants given under the Clean Water Act on water pollution down stream for plant i. The problem is same in case of ATE. At the same time, for a plant i, we cannot observe both outcomes. Hence it is impossible to measure ATE.

How would a realistic experiment look like?

An RCT where the treatment is assigned to provinces randomly. Then we can calculate the effect of treatment i.e grants given to plant i on water pollution. When the treatment assigned randomly we can say that there will not be any selection problem by design.This means that $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$ and $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$ which transforms our above equation into

$tau^{ATE} = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$

Then the ATE will be equal to Naive estimator $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$

For this to workout, we assume that the outcome is solely affected by the treatment and there is 100% compliance and there are no spillover effects among the treated or control groups.

From the above the Naive estimator τ_N is given by $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$. The naive estimator is determined from sample averages. As we cannot observe both outcomes for i at a same time, here with the Naive estimator we observe $Y_i(D_i = 1)$ and $Y_j(D_j = 0)$, where i is not equal to j. For this to workout , we assume that the unconditional expectation of outcome is same as its conditional outcome. In math, this is given by $E[Y_i(1)] = E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$ and $E[Y_i(0)] = E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$.

Here we need to observe how different are plants that are given grants are to the plants that arent given grants under the Clean water act. i.e how different is the treatment group to the control group.

Three examples

1) Selection of observables Consider the case where the treatments plants which are given grants are very different from the treatment plants that are not given grants. An example of the same being the treatment plant being with discharge from a city and the other plant with the discharge from a village. Lets say that the plants similar to first are under the treatment group and the plants similar to the second are under the control group. In this case, the water pollution levels downstream for the plants similar to first will be very high compared to the water pollution levels downstreat for the plants like the second. In such cases, the Naive Estimator will underestimate the ATE of the Clean Water act on water pollution levels down the stream.

$\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$

2) Selection of unobservables Lets say that the plants in US where the study is being conducted have the same water pollution levels. Lets say that the plants that comes under the treatmeent group has the governments that are rich in knowledge reduce the water pollution levels more compared to the plants that comes under the control group. In such cases, the Naive Estimator will underestimate the ATE of the Clean Water act on water pollution levels down the stream.

$\tau_N = \bar{Y}(D=1) - \bar{Y}(D=0)$

A similar case can be constructed where the Naive Estimator overestimates the ATE.

3) Non compliance Lets consider the plants that are under control group has human colonies that are educated and understand the problems of waste water. The people of these colonies will be more pasionate about solving the water problem even without the grants given to the treatment plant. In such cases, they being more passionate about the water issue will find ways to reduce the water pollution levels down the stream. When such cases, the Naive estimator cannot accurately estimate the ATE of the treatment i.e the 1972 Clean Water Act.

$\tau_N = \bar{Y}(D=1) - \bar{Y}(D=0)$

4) Spill Overs Spill overs is a case where there is overflow of the effect into other places. A classic case of the spill overs is when the water from downstream of one plant flows through to the water of downstream of another plant. Now in the case that the first plant is treated i.e given grants under the Clean Water act and the second plant is under the control group, this causes a problem where the treatment of the first plant spills over into the other plant and changes the water pollution levels down the stream of second plant. In such cases, the Naive estimator $\tau_N = \bar{Y}(D=1) - \bar{Y}(D=0)$ produces an inaccurate estimate of the ATE of the treatment i.e the 1972 Clean Water Act in US.

Question IB

Copy down the authors' main regression specification (and be sure to list which equation number this is in the paper). Explain, in words and math, what treatment parameter the authors are recovering with their main specification. Does this approach recover the population average treatment effect? If yes, why? If no, why not? What assumptions are required for this regression to recover the causal effect of interest? Do you think these assumptions are likely to be satisfied in this context? Why or why not? Include references to evidence presented in the paper to support your conclusion.

authors' main regression specification is the equation 3, given below

$Q_{pdy} = \gamma.G_{py}.d_d + X'_{pdy}.\beta + \eta_{pd} + \eta_{py} + \eta_{dwy} + \epsilon_{pdy}$

This is a DDD model containing two observations i.e mean water quality upstream denoted by d = 0 and mean water quality downstream denoted by d = 1 for a treatment plant p and in year y

Terms (left to right): Q_{pdy}: Water quality Q for a plant p in year y, categorized by upstream d = 0 or downstream d = 1 γ: The ATE estimate of each grant on downstream water pollution
G_{py}: Total number of grants given to a plant p till the year y
X'_{pdy}: Temperature and precipitation control matrix, for plant p in year y, contains d where d = 1 is downstream and d = 0 is upstream η_{pd}: Fixed effects that contain control for the time-invariant characteristics that are specific with d, categorized by d where d = 1 is downstream and d = 0 is upstream η_{py}: The fixed effects that affects the outcome specific to the year y for a plant p. These characteristics affects both the streams denoted by d. One example of this characteristic is change in water pollution due to increased activity around the plants. η_{dwy}: The fixed effects encompassing d (downstream/upstream), w(basin) and y(year) that affect the water quality. ϵ_{pdy}: Error Term

Does this approach recover the population average treatment effect? If yes, why? If no, why not?What assumptions are required for this regression to recover the causal effect of interest? Do you think these assumptions are likely to be satisfied in this context? Why or why not? Include references to evidence presented in the paper to support your conclusion.

As we cannot implement an RCT approach for randomization, we need to undertake a grouping comparision based approach and compare pre treatment vs post treatment, treatment group vs control group etc. The

DID estimator does recover the population ATE for each grant given, on the water pollution level down stream of plant p.This can be grouped to comparision with the water pollution levels upstream the plant p. Directly from the lecture notes, the underlying assumption for DDD approach is that the differences in trends between treated and untreated impacts the affected group and the unaffected group in the same way. This means that the trend differences in the plants that were given grants and the plants that arent given grants affect both the downstream water pollution levels and upstream water pollutions levels in the same way. By this assumptions the concerns of Coincident treatments, Non parallel trends in treated/control groups and the Non parallel counterfactual trends can be adressed through DDD. The underlying assumption mentioned above and that the $G_{py}.d_d$ is independent of error term ϵ_{pdy} given other independent variables are required for the regression equation to recover a unbiased estimate of γ. This would reiterate addressing the concerns mentioned earlier and result in a treatment that satisfies the randomized assignment of treatment. The assumption regarding the parameter $G_{py}.d_d$ would not satisfy if there is any dependency observed with unobservable characteristics that affect our determined outcomes i.e the water pollution levels. This means a trend in the pre treatment period that is unaccounted for. The authors however discussed(through an event study design in Equation 4 and shown in Figure III) in the paper that years before grants were given, they observed the parameters to be statistically significant and exhibit no trend in the pre treatment period. Hence I believe that the assumptions mentioned above would not be violated.

Question IC

Describe the main results of the paper. Include a discussion of (at a minimum) one table and one figure, in which you interpret the estimated coefficients and describe their magnitudes. What is the main policy take-away of the paper?

Main results of the paper is effect of grants given under the Clean Water Act on the water pollution levels. The paper also discusses the effectiveness of the program, water pollution trends.

EFFECTS OF CLEAN WATER ACT GRANTS ON WATER POLLUTION (Table II Page 378, Figure III Page 377): Dissolved oxygen deficits decrease by 0.7 percentage points and the probability that downstream water is not fishable decreases by 0.7 percentage points with each grant given. The other pollutants decrease as well—BOD falls by about 2.4%, fecal coliforms fall by 3.6%, and the probability that downstream waters are not swimmable by about half a percentage point. The point estimate implies that each grant decreases Total suspended solids by 1%, though this is imprecise. The magnitue is small and the coefficients are indistinguishable from zero as shown in the Figure III. Event study graphs corresponding to equation (4) support these results. In years before a grant, the coefficients are statistically indistinguishable from zero, have modest magnitude, and have no clear trend (Figure III). This implies that pollution levels in upstream and downstream waters had similar trends before grants were received. In the years after a grant, downstream wa- ters have 1—2% lower dissolved oxygen deficits, and become 1—2% less likely to violate fishing standards. These effects grow in magnitude over the first 10 years, are statistically significant in this period, and remain negative for about 30 years after a grant. This means that the water pollution levels in the water upstream and water downstream have same trends years before the grants were given. The same is reiterated by the event study design that was conducted years before the grant is given. The same is determined by the the Equation 4 in paper.

EFFECTS OF CLEAN WATER ACT GRANTS ON HOUSING DEMAND (Table V page 384):

Table V analyzes how Clean Water Act grants affect housing. Column (1) shows estimates for homes within a quarter mile of downstream waters. Column (2) adds controls for dwelling characteristics, and for baseline covariates interacted with year fixed effects. Column (3) include all homes within 1 mile, and column (4) includes homes within 25 miles.

Panel A reports estimates of how grants affect log mean home values. The positive coefficients in the richer specifications of columns (2) through (4) are consistent with increases in home values, though most are statistically insignificant. Column (4) implies that each grant increases mean home values within 25 miles of affected waters by 0.024 percentage points. The 0.25- or 1.0-mile estimates are slightly larger, which is consistent withthe idea that residents nearer to the river benefit more from water quality.

Panel B analyzes how grants affect log mean rental values. These estimates are even less positive than the estimates for housing. The estimate in column (4), including homes within a 25-mile radius of downstream rivers, is small and statistically insignificant but actually negative.

Panels A and B reflect the classic hedonic model, with fixed housing stock.

Panel C estimates the effect of grants on log housing units and Panel D on the log of the total value of the housing stock. They suggest similar conclusions as Panels A and B. Most of these estimates are small and actually negative.

Overall, the panel A shows evidens that in the years after a plant received a grant, the value of homes within 0.25 miles of the downstream river increase. However, the Panel B doesnt show any evidence that the homes within 25 miles of the downstream river increase after the plant is given a grant.

Main policy take-away:

Though the U.S. water pollution has declined since 1972 that encompasses $650 billion in expenditure, some evidence suggests it may have declined faster before 1972. Each grant significantly decreased pollution for 25 miles downstream, and these benefits last for around 30 years. The share of waters that are fishable has grown by 12 percentage points since the Clean Water Act.The point estimates imply that the benefits of the Clean Water Act's municipal grants exceed their costs if these unmeasured components of willingness to pay are three or more times the components of willingness to pay that we measure. There is evidence that the net benefits of Clean Water Act grants vary over space in tandem with population density and the popularity of water-based recreation. The estimated ratio of the change in housing costs to total grant costs may provide a lower bound on the true benefit/cost ratio of this grant program because we abstract from nonuse ("existence") values, general equilibrium effects, potential changes in sewer fees, and the roughly 5% longest recreational trips. Though estimates of increases in housing values are generally smaller than costs of the grant projects, there is almost no evidence that local residents value these grants.

A comparision study with Air pollution resulted in observing similar trends and patterns. Thus following the evidence from the paper that there is evidence that net benefits of Clean water act grants vary over space in tandem with the population density and the popularity of water-based recreation, the pollution regulations be in air or water can create net benefit on a large scale and produce higher social welfare goes hand in hand when allowed to vary over space.

```
library(knitr)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

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

```
library(haven)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.0.1

## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 4.1.2
```

```
##
## 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(broom)

library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 4.1.3
```

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library("ivreg")
```

```
## Warning: package 'ivreg' was built under R version 4.1.2
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.2

##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:purrr':
##
##     some

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(AER)
```

```
## Warning: package 'AER' was built under R version 4.1.3

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 4.1.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.1.2

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Warning: package 'sandwich' was built under R version 4.1.2

## Loading required package: survival

## Registered S3 methods overwritten by 'AER':
##   method              from
##   print.ivreg         ivreg
##   print.summary.ivreg ivreg
##   summary.ivreg       ivreg
##   vcov.ivreg          ivreg
##   bread.ivreg         ivreg
##   estfun.ivreg        ivreg
##   hatvalues.ivreg     ivreg
##   predict.ivreg       ivreg
##   anova.ivreg         ivreg
##   terms.ivreg         ivreg
##   model.matrix.ivreg  ivreg
##   update.ivreg        ivreg

##
## Attaching package: 'AER'
```

```
## The following objects are masked from 'package:ivreg':
##
##     ivreg, ivreg.fit
```

```
library(stargazer)

library(lfe)
```

```
## Warning: package 'lfe' was built under R version 4.1.3
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## Attaching package: 'lfe'
```

```
## The following object is masked from 'package:lmtest':
##
##     waldtest
```

```
library(plm)
```

```
## Warning: package 'plm' was built under R version 4.1.3
```

```
##
## Attaching package: 'plm'
```

```
## The following object is masked from 'package:lfe':
##
##     sargan
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

```
library(ggplot2)
```

Question II

A local activist group, the Universal Committee to Heighten Interest in Cars Against Greenhouse Outlays (UCHICAGO) promotes the adoption of electric vehicles in Chicago. They would like you to help them design a pilot program to demonstrate the impacts of a new electric vehicle (EV) subsidy they are trying to roll out at scale.

Question II A

UCHICAGO hypothesizes that providing homeowners with subsidies for electric vehicles will increase miles driven in EVs. Using the potential outcomes framework, describe the impact of treatment - a 20% EV subsidy – on electric vehicle adoption, household electricity use, and electric vehicle driving at the household level. EV adoption is a binary variable; household electricity use is measured in kWh, and electric vehicle driving is measured in eVMT. Explain to UCHICAGO what the ideal experiment would be for answering this question. Describe the dataset you'd like to have to carry out this ideal experiment and use math, words, and the potential outcomes framework to explain what you would estimate and how you would do so. Make sure to be clear about the unit of analysis (ie, what is i here?)

Ideal Experiment: For the same household i, we need to observe the following

Electric vehicle adoption: 1) Probability of Electric Vehicle adoption when EV subsidy is given - i.e treatment 2) Probability of Electric Vehicle adoption when EV subsidy is not given i.e control

Difference between 1 and 2 gives us the effect of EV Subsidy on Electric Vehicle adoption for household i.

Household electricity use: 3) Total household electricity use(kWh) when EV subsidy is given - i.e treatment 4) Total household electricity use(kWh) when EV subsidy is not given - i.e control

Difference between 3 and 4 gives us the effect of EV Subsidy on Total household electricity use(kWh) for household i.

Electric Vehicle driving 5) Total electric vehicle driving(eVMT) when EV subsidy is given - i.e treatment 6) Total electric vehicle driving(eVMT) when EV subsity is not given - i.e control

Difference between 3 and 4 gives us the effect of EV Subsidy on Total electric vehicle driving(eVMT) for household i.

Potential outcomes Framework:

Let i be the individual household Treatment indicator D_i Treated: $D_i = 1$: EV susidy given to household i Untreated: $D_i = 0$: EV subsidy not given to household i

Electric Vehicle adoption: Outcome treated: $Y_i(D_i = 1)$: Probability of Electric Vehicle adoption for household i, when EV subsidy is given - Treatment Outcome untreated: $Y_i(D_i = 0)$: Probability of Electric Vehicle adoption for household i, when EV subsidy is not given - Control

Household electricity use(kWh): Outcome treated: $Y_i(D_i = 1)$: Total Household electricity use(kWh) of household i, when EV subsidy is given - Treatment Outcome untreated: $Y_i(D_i = 0)$: Total Household electricity use(kWh) of household i, when EV subsidy is not given - Control

Electric Vehicle driving(eVMT): Outcome treated: $Y_i(D_i = 1)$: Total Electric Vehicle driving(eVMT) by household i, when EV subsidy is given - Treatment Outcome untreated: $Y_i(D_i = 0)$: Total Electric Vehicle driving(eVMT) by household i, when EV subsidy is not given - Control

We get the impact of treatment(i.e EV Subsidy) τ_i from the difference between the above outcomes i.e for the outcome variables Electric Vehicle adoption, Household electricity use(kWh) and Electric Vehicle driving(eVMT):

Eqn(1) $\tau_i = Y_i(D_i = 1) - Y_i(D_i = 0)$

While for an outcome variable i.e either of Electric Vehicle adoption, Household electricity use(kWh), Electric Vehicle driving(eVMT) we need $Y_i(D_i = 1)$ and $Y_i(D_i = 0)$ at a given time to compute the impact of treatment, the problem is that at a given time, we cannot observe both the outcomes $ or $ we can only observe either $Y_i(D_i = 1)$ or $ Y_i(D_i = 0)$ in real world.

In case a household is treated (i.e EV Subsidy is given), then the observed outcome would be $Y_i(D_i = 1)$, and $Y_i(D_i = 0)$ would become an unobserved outcome. Due to the un-observable outcome *or* not being able to observe both the outcomes at a given time, measuring τ_i is impossible.

Average Treatment Effect $ tauˆ{ATE}$

- $ tauˆATE = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$

ATE measures the average effect of treatment across a population of households. ATE measures the effect of EV Subsidy on the outcome variables Electric Vehicle adoption, Household electricity use(kWh), Electric Vehicle driving(eVMT). The problem is same in case of ATE. At the same time, for a household i, we cannot observe both outcomes for each outcome variable. Hence it is impossible to measure ATE.

As Ideal experiment is not possible to measure ATE, we frame a realisitc experiment as follows.

Realistic Experiment:

An RCT where the treatment is assigned to provinces randomly. Then we can calculate the effect of treatment i.e EV Subsidy on each of the outcome variable Electric Vehicle adoption, Household electricity use(kWh), Electric Vehicle driving(eVMT). When the treatment assigned randomly and the distribution of the observables and the unobservables are same across the treated and untreated, we can take that there is no selection problem by design. And that the D_i is exogenous.

Hence we get, $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$ and $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$

and $tau^{ATE} = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$

Then the ATE will be equal to Naive estimator $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$

For this to workout, we assume that the outcome is solely affected by the treatment and there is 100% compliance (for all i, $R_i = D_i$) and there are no spillover effects among the treated or control groups.

Question II B

UCHICAGO have secured funding to run a randomized trial to test their hypothesis. However, their funder is worried about implementing an individually-randomized design. In particular, they are concerned that control group individuals may be less likely to purchase an EV if all of the cars on the market are bought by the treatment group. Is this a problem for an RCT that is randomized at the individual level? If yes, explain why, and describe what this would do to your treatment effects relative to the truth. If no, explain why not. Assume for the remainder of Question II that everyone offered a subsidy by UCHICAGO purchases an EV.

Given that UCHICAGO's funder are in particular concerned that the control group individuals may be less likely to purchase an EV if all of the cars on the market are bought by the treatment group. This is the case of Spill over effects. We can understand this is a way that by giving EV Subsidy for the treatment group and thus all cars are bought by the treatment group, it can negatively impact the purchasing of an EV on the control group. This would lead to an over estimation of the treatment on the outcome. Shown as follows:

As the outcome value $E[Y_i(D_i = 1)]$ increases i.e the Electriic Vehicle adoption in treatment group increases with the treatment EV Subsidy, the outcome value $E[Y_i(D_i = 1)]$ decreses i.e the Electric Vehicle adoption in control group decreases with the treatment EV subsidy. Hence the overestimation happens as the ATE i.e τ^{ATE} is given by $\tau^{ATE} = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$.

Due to this, the SUTVA i.e Stable Unit Treatment Variable Assumption which implies "no unmodeled spillovers" that is required for RCT's will not be satisfied anymore.

Question II C

The funder is adamant that they will not support an individually-randomized design. They would like UCHICAGO to instead provide subsidies at the neighborhood level. Will this address the concern they raised in IIA? If yes, explain how. If no, explain why this will be no better than the individually-randomized design.

UCHICAGO to provide subsidies at neighbourhood level.

The concern raised in IIB, the spill over effects will be adressed by providing subsidies at the neighbourhood level. This is by implementing a Radomized Saturation Design by considering each neighbourhood as a cluster. Can be done by Randomizing neighborhood clusters into treatment intensities to compare high and low intenisity places. And by randomizing units within a neighborhood cluster to compare treatment and control units.

For neighborhood c, let $\pi_c \in [0, 1]$ be the neighborhood treatment saturation and within a neighborhood c let D_{ic} be the treatment status for a unit i

Here $D_{ic} = 1$ means a treated individual, $D_{ic} = 0$ and $\pi_c = 0$ means pure control, and $D_{ic} = 0$ and $\pi_c > 0$ means within a neighborhood control. The outcome Y_{ic} depends on D_{ic} and D_{jc}

$Y_{ic} = f(D_{ic}, D_{jc}; X_{ic}, \epsilon_{ic})$

D_{ic} treatment status for unit i in neighborhood c D_{jc} treatment status for unit j in neighborhood c

Assumptions:

$Y_{ic} \perp D_{jd}$ for all d not equal to c. We assume that the household i's outcome in neighbourhood c is not affected by households in other neighborhoods. This means we assume there is no interference across the clusters.

Treatment parameters under Randomized Saturation Design:

Intent to treat i.e ITT i.e difference between those offered treatment and pure controls given by $\tau^{ITT}(\pi)$ "Spillover on the non-treated" i.e SNT i.e difference between control units in treated clusters and pure controls givent by $\tau^{SNT}(\pi)$ and the "Spillover on Treated" i.e saturation dependent spillover effect given by $\tau^{ST}(\pi)$. Total causal effect i.e overall difference between treated and control clusters given by $\tau^{TCE}(\pi)$ and the "Treatment on Uniqely treated" i.e where we treat only on eunit with no spillover given by $\tau^{TUT}(\pi)$

We get $\tau^{ITT}(\pi) = \tau^{ATE}(\pi)$.

From the above , we can say that the Randomized saturation design and the regular RCT differs where the Randomized saturation design allows spillovers or the violation for the SUTVA assumption. Thus the concern raised in IIB, the spill over effects will be adressed by providing subsidies at the neighbourhood level by considering them as clusters through a Randomized Saturation design.

Question II D

After listening to the funder's feedback, UCHICAGO has decided that they are actually interested in studying how providing EV subsidies to some people in a neighborhood impacts EV adoption, home electricity use, and EV driving for people who don't get the subsidy. The funder is excited about this too, and is willing to devote substantial funding to the project. Given a large budget, describe an RCT design that will allow you to measure treatment effects on both households who get a subsidy from UCHICAGO and households who don't. Make sure to describe any necessary steps, and clearly lay out any treatment arms. A tree diagram may be helpful. Use words and math to explain what treatment parameters you can estimate with this design. Be sure to be clear about the comparisons you are making. Write down a regression equation that you would use to estimate these treatment parameters, and describe how you would interpret any coefficients you recover. Finally, describe how you would use your estimates to recommend to UCHICAGO whether they should scale their subsidy program or not.

describe an RCT design that will allow you to measure treatment effects on both households who get a subsidy from UCHICAGO and households who don't

Randomized saturation design:

Start with N households in C (disjoint) neighborhoods (clusters) Step 1: Randomly assign neighborhoods (clusters) a treatment saturation $\pi_c \in [0, 1]$ Choose a set of pre-determined saturation and also have a pure control neighborhood (cluster) with $\pi_c = 0$ i.e no households in the neighborhood were given treatment i.e EV subsidy.

Step 2: In each cluster, randomly assign $\pi_c.N_c$ units into treatment Now D_{ic} is the treatment status for unit i in cluster c

This Randomized saturation design results in three types of units: Treated individual: $D_{ic} = 1$ as usual Share in sample: Mu Pure control: $D_{ic} = 0$ and $\pi_c = 0$ Share in sample: fi Within-neighborhood(within cluster) control: $D_{ic} = 0$ and $\pi_c > 0$

Y_{ic} denotes the outcome variable (EV adoption/ home electricity use/ EV driving) for household i in cluster c

A tree diagram may be helpful.

High intensity is High saturation neighborhoods Low intensity depicts Low saturation neighborhoods And there will be another tree that depicts "No Saturation neighborhoods" with only control block meaning pure control.

Use words and math to explain what treatment parameters you can estimate with this design. Be sure to be clear about the comparisons you are making.

THe Randomized Saturation design allows Stable Unit Treatment Value Assumption (SUTVA) violations. RS Designs open the door to new treatment parameters:

Intent to treat (ITT): Difference between those offered treatment and pure controls, saturation level $\pi_c = \pi$ Given by $\tau^{ITT}(\pi) = E[Y_{ic} \mid D_{ic} = 1, \pi_c = \pi] - E[Y_{ic} \mid D_{ic} = 0, \pi_c = 0]$ We can estimate neighborhood ATE $\tau^{ATE}(\pi)$ using the Intent to Treat (ITT) $\tau^{ITT}(\pi)$ for saturation level π_c i.e $\tau^{ATE}(\pi) = \tau^{ITT}(\pi)$

Spillover on the non-treated (SNT): $\tau^{SNT}(\pi)$: Difference between control units in treated clusters and pure controls, saturation level $\pi_c = \pi$ Given by $\tau^{SNT}(\pi) = E[Y_{ic} \mid D_{ic} = 0, \pi_c = \pi] - E[Y_{ic} \mid D_{ic} = 0, \pi_c = 0]$ The SNT $\tau^{SNT}(\pi)$ gives a good estimate of how treatment i.e prividing EV subsidy to some households in neighborhood (cluster) c impacts the outcome variables for people who are not given EV subsidy

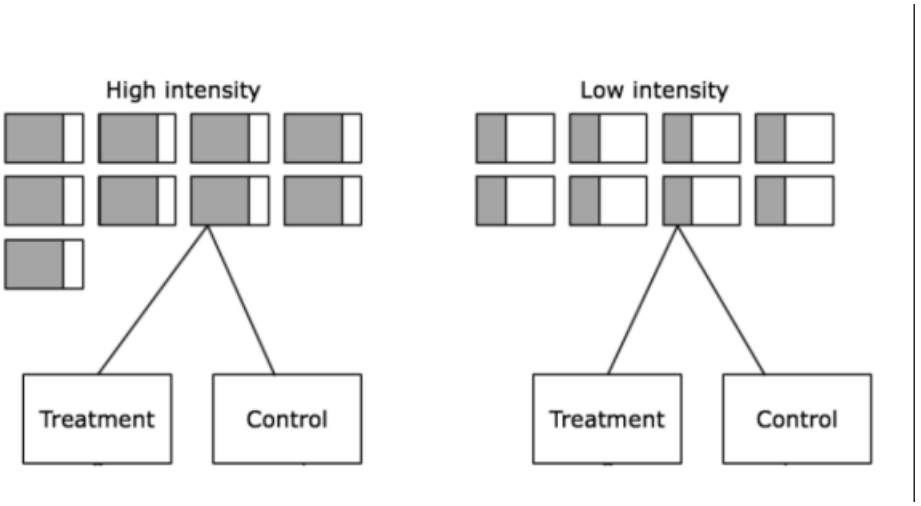


Figure 1: Tree diagram - High and low saturation

Total causal effect (TCE): $\tau^{TCE}(\pi)$: Overall cluster difference between treated and control clusters, saturation level $\pi_c = \pi$ Given by $\tau^{TCE}(\pi) = E[Y_{ic}|\pi_c = \pi] \ - \ E[Y_{ic}|\pi_c = 0] = \pi.\tau^{ITT}(\pi) + (1-\pi).\tau^{SNT}(\pi)$ In other words $D_{ic} = 1, \pi_c > 0$ gets tau^{ITT} and $D_{ic} = 0, \pi_c > 0$ units get tau^{SNT}

Treatment on the Uniquely Treated (TUT): $\tau^{TUT}(\pi)$: This is the ITT were we to only treat one unit (no spillovers!)
Given by $\tau^{TUT} = E[Y_{ic}|D_{ic} = 1, \pi_c = 0] \ - \ E[Y_{ic}|D_{ic} = 0, \pi_c = 0] = \tau^{ITT}(\pi = 0)$ This is not a function of π

Spillover on the Treated (ST): $\tau^{ST}(\pi)$: This is the saturation-dependent spillover effect (only spillovers!)
Given by $\tau^{ST}(\pi) = E[Y_{ic}|D_{ic} = 1, \pi_c = \pi] \ - \ E[Y_{ic}|D_{ic} = 1, \pi_c = 0]$

Write down a regression equation that you would use to estimate these treatment parameters, and describe how you would interpret any coefficients you recover

Regression equation

$Y_{ic} = \alpha \ + \ \sum_{\pi \neq 0} \tau^{trt}.D_{ic}.1[\pi_c = \pi] + + \ \sum_{\pi \neq 0} \tau^{ctrl}.S_{ic}.1[\pi_c = \pi] + \epsilon_{ic}$

Y_{ic}: Outcome of household i in neighborhood (cluster) c
$D_{ic}.1[\pi_c = \pi]$: Indicator for a treated unit, neighborhood(cluster) saturation π_c
$S_{ic}.1[\pi_c = \pi]$: Indicator for a control unit, neighborhood(cluster) saturation π_c
ϵ_{ic}: Error term

These compared with the pure controls i.e $\pi_c = 0$ results in the parameters as follwos:

Intent to treat (ITT): $\tau^{ITT}(\pi)$: Gives $\hat{\tau}^{ITT}(\pi) = \hat{\tau}^{trt}$

Spillover on the non-treated (SNT): $\tau^{SNT}(\pi)$: Gives $\hat{\tau}^{SNT}(\pi) = \hat{\tau}^{ctrl}$

Total causal effect (TCE) $\tau^{TCE}(\pi)$: Gives $\hat{\tau}^{TCE}(\pi) = \pi.\tau^{ITT}(\pi) + (1-\pi).\tau^{SNT}(\pi) = \pi.\hat{\tau}^{trt}(\pi) + (1-\pi).\hat{\tau}^{ctrl}(\pi)$

Finally, describe how you would use your estimates to recommend to UCHICAGO whether they should scale their subsidy program or not.

First by performing the Randomized saturation design and determine the following parameters: Intent to treat (ITT) $\tau^{ITT}(\pi)$ and Spillover on the non-treated (SNT) $\tau^{SNT}(\pi)$ and Total causal effect (TCE) $\tau^{TCE}(\pi)$

We use the above three terms to mae recommend to UCHICAGO whether they should scale their subsidy program or not, we do it as follows

Intent to treat (ITT) $\tau^{ITT}(\pi)$: If ITT for outcome Electric Vehicle adoption is positive and statistically significant, that means that the Electric Vehicle adoption is greater in households for whcih EV subsidy is given.

Spillover on the non-treated (SNT) $\tau^{SNT}(\pi)$: If SNT for outcome Electric Vehicle adoption is positive and statistically significant, that means EV subsidy for a few households in the neighborhood shows positive impact on the households who arent given subsidy. Similarly, if SNT is negative for outcome Electric Vehicle adoption, and is statistically significant, that means that EV subsidy for a few households in the neighborhood negatively affects the EV adoption on the households who arent given subsidy.

The Total causal effect (TCE) $\tau^{TCE}(\pi)$: This terms will help us determine the overall difference at a neighborhood level, i.e the overall difference between treated neighborhoods and control neighborhoods.

Question III

Around the world, women are under-represented in politics. A bunch of people have hypothesized that having more women in positions of political power could improve well-being. A policy advocacy group, Powerful, Respected, and Organized Government Requires that All Men Elect Valuable Accomplished Ladies (PROGRAMEVAL), is interested in demonstrating the impact of women in government on local public goods provision in India. They've asked you for help with their analysis.

Question III A

PROGRAMEVAL would like you to compare the average number of public goods (roads, schools, public buildings, et cetera) in towns with female-headed governments as compared with towns that have male-headed governments. Describe this comparison in math and words. Under what conditions would this comparison estimate the causal effect of female leaders on public goods provision? Provide two concrete examples of reasons why this comparison may be problematic.

Let i be the individual town in India where $i \in \{1, 2, ...N\}$. Treatment indicator D_i where $D_i \in \{0, 1\}$
Treated: $D_i = 1$: a female headed government
Untreated: $D_i = 0$: not a female headed government

Outcome treated: $Y_i(D_i = 1)$: Total number of public goods (roads, schools, public buildings, et cetera) in town i when a female-headed government - Treatment Outcome untreated: $Y_i(D_i = 0)$: Total number of public goods (roads, schools, public buildings, et cetera) in town i when not a female-headed government - Control

PROGRAMEVAL would like to compare the average number of public goods (roads, schools, public buildings, et cetera) in towns with female-headed governments as compared with towns that have male-headed governments.

We get the impact of treatment τ_i from the difference between the above outcomes.

$\tau_i = Y_i(D_i = 1) - Y_i(D_i = 0)$

While we need both the outcomes at a given time to compute the impact of treatment, the problem is that at a given time, we can only observe either \$Y_i(D_i = 1) \$ or \$ Y_i(D_i = 0)\$.Thus measuring τ_i is impossible.

Average Treatment Effect \$ tauˆ{ATE}\$

$tau^ATE = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$

ATE measures the average effect of treatment across a population of towns. The problem is same in case of ATE. At the same time, for a province i, we cannot observe both outcomes. Hence it is impossible to measure ATE.

How would a realistic experiment look like?

An RCT where the treatment is assigned to twons randomly. Then we can calculate the effect of treatment. When the treatment assigned randomly we can say that there will not be any selection problem by design.This means that $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$ and $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$ which transforms our above equation into

$tau^{ATE} = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$

Then the ATE will be equal to Naive estimator $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$

For this to workout, we assume that the outcome is solely affected by the treatment and there is 100% compliance and there are no spillover effects among the treated or control groups.

From the above the Naive estimator τ_N is given by $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$. The naive estimator is determined from sample averages. As we cannot observe both outcomes for i at a same time, here with the Naive estimator we observe $Y_i(D_i = 1)$ and $Y_j(D_j = 0)$, where i is not equal to j. For this to workout , we assume that the unconditional expectation of outcome is same as its conditional outcome. In math, this is given by $E[Y_i(1)] = E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$ and $E[Y_i(0)] = E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$.

For this to workout, we assume that the outcome is solely affected by the treatment and there is 100% compliance and there are no spillover effects among the treated or control groups.

From the above the Naive estimator τ_N is given by $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$. The naive estimator is determined from sample averages. As we cannot observe both outcomes for i at a same time, here with the Naive estimator we observe $Y_i(D_i = 1)$ and $Y_j(D_j = 0)$, where i is not equal to j. For this to workout , we assume that the unconditional expectation of outcome is same as its conditional outcome. In math, this is given by $E[Y_i(1)] = E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$ and $E[Y_i(0)] = E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$.
The issues with naive estimator are as follows:

Example: Selection of observables Consider a case where there are rich towns or with rich governments and the poor towns with just enough resources. Lets say that rich towns have a female-headed governments i.e they fall under the treatment group and the poor towns have a male-headed governments i.e they fall under the control group. In such cases, the Public Goods number for the rich towns ie the treatment group will be comapritively higher than the Public goods number in the poor towns i.e the control group. In such cases, the Naive estimator $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$ will overestimate the ATE of treatment i.e having a female-headed governments on the number of Public goods i.e the outcomes.

A similar case can be constructed where the Naive estimator will underestimate the ATE.

Example: Selection on observables: Now lets assume that the towns with the female-headed government i.e treatment group and the towns with male-headed government ie the control group have the same number of public goods. Now lets say that the people in the towns with female-headed government are very educated and the towns has organized economy. And the towns with male-headed governments i.e the control group has a majority unorganized economy. Though these towns have same number of public goods, the organized economy gives alot back to the government in form of taxes when compared to the towns where there is majority unorganized economy. This characteristic affects the outcomes in question. In such cases the Naive estimator $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$ will over estimate the ATE of treatment i.e having female-headed government on the outcome i.e the number of public goods.

Question III B

PROGRAMEVAL gets it - this is not the best comparison. However, they have data on a bunch of other town characteristics: per-capita income, number of residents, year of incorporation, average population age, and share of gross city product devoted to manufacturing. Describe, using math and words, a comparison between female- and male-headed towns which leverages these administrative data. Under what conditions would this comparison estimate the causal effect of female leadership on public goods provision? Provide two concrete examples of reasons why this comparison may be problematic (different from what you described above).

Characteristics: per-capita income, number of residents, year of incorporation, average population age, and share of gross city product devoted to manufacturing.

We can take an SOO approach with the above characteristics in the data as covariates.

Assumptions:

Common Support Assumption For all the possible covariate X's, when we have a large sample we should be able to observe both the treated and untreated as we have a significantly large sample. As we can observe both the treated and untreated, treatment effects can be estimated. $0 < Pr(D_i|X = x^0) < 1, \forall x^0$ In our context, the covariates are per-capita income, number of residents, year of incorporation, average population age etc.. and treated means female-headed town and untreated means male-headed towns. We can then calculate the ATE by determining the τ^{SOO} i.e the weighted average across all covariates.

$\int \tau^{SOO}.dP(X) = \int (E[Y_i(1) \mid Xi = x] - E[Y_i(0) \mid Xi = x]).dP(X)$

This assumption can be validated.

Conditional Independence assumption: When conditioned on the X_i's, the potential outcomes of a unit are orthogonal to the treatment. In this context, when a given X_i i.e per-capita income, number of residents, year of incorporation, average population age etc are independed of treatement i.e having a female-headed government. This assumptions gives a safe base for comparision of estimates with other units. This assumptions helps us to control group counterfactuals are good estimates of the treatment group counterfactuals. As we work with potential outcomes and not observed, we cannot check for validity of this assumption.

$(Y_{1,i}, Y_{0,i}) \perp D_i|X$

ATE (Average Treatment Effect) τ^{SOO}:

$\tau^{SOO} = E[Y_i(1)|X_i = x] \ - \ E[Y_i(0)|X_i = x]$

ATE $= \int \tau^{SOO}.dP(X) = \int (E[Y_i(1) \mid Xi = x] - E[Y_i(0) \mid Xi = x]).dP(X)$

As said above, we cannot check for validity of the conditional independence assumption.

Example 1: Curse of Dimensionality: The dataset size increases with the number of covariates and for the first assumption to hold we need a very large data set that may not be feasible always. In a case where the assumption doesnt hold, we cannot use the SOO design to accurately determine the τ^{SOO} to estimate ATE. In this context the example where the assumption might not hold true is the case where the matched groups has only either female-headed governemnts or male-headed governemnts but not both which leads to this assumption to not be satisfied thereby producing an inaccurate estimate of ATE.

Example 2: Conditional Independence Assumption Lets say we have a variable that was ommitted such as the conservativeness of the town. Such characteristics affects the random assignment of treatment and the outcomes and thereby the treatment can no longer be assumed as random eventhough all the covariates are controlled for. This is due to the ommitted variable which results in a bias and produces a inaccurate estimator τ^{SOO} for estimating the ATE.

Question III C

PROGRAMEVAL understands your concerns, but has some in-house machine learning experts. They tell you that they can use this same administrative data to solve your issues. Do you agree? Why or why not? Be specific.

Machine learning can be helpful in understanding the relation that Covariates maintain with the potential outcomes and with the treatment. We can use machine learning to determine the effect of covariates on outcomes and effect of covariates on the treatment and filter out the covariates that are not important in the design. With only picking few covariates and primarily the ones that are very important and including them in the analysis will help deal with the case of Curse of Dimensionality mentioned in the question III B.

While Machine Learning can help us mitigate the issues of Curse of Dimentionality and provide us with important covariates for the design, Machine Learning cannot be used to tackle the other problems mentioned earlier such as the Conditional Independence Assumption and the Common Support Assumption. The problems that arise with these are due to unobserved covariates like the "Conservativeness of town" mentioned in IIIB. Such unobserved or ommitted covariates cannot be dealt with Machine Learning and hence Conditional Independence Assumption will not be satisfied. In the Common Support Assumption case, as mentioned earlier, Machine Learning cannot solve the problem of not observing both the treated and untreated together in the matched group.

Thus we can say that assuming the basic SOO assumptions, namely Conditional Independence Assumption and the Common Support Assumption are satisfied, Machine Learning can be used in improving the research approach.

Question III D

PROGRAMEVAL forgot to tell you that, in India, certain local government positions are "reserved" for women – meaning only women can run for office to fill these seats (this is, again, a Real Thing!). They inform you that towns are selected to have reserved seats based on their political party. In Uttar Pradesh, an Indian state, all towns are put on a list, ordered by the share of women in the population. Each election cycle, the top 500 towns on the list are required to reserve the leadership positions for women (though, be warned – official rules aren't always perfectly followed). PROGRAMEVAL asks you whether you can use Uttar Pradesh as a test case. Describe, in math and words, the research design you would use to leverage this new information. Be sure to include a regression equation. Under what conditions would this approach estimate the causal effect of female leadership on public goods provision? For whom is this causal effect identified?

We implement a Regression Discontinuity design. We use the variable rank that includes a threshold of 500 to determine treatment. We can thus under RDD, randomly assign by observing outcomes of towns just above and below the threshold to determine the assignment.

Here Rank is rank in the list of towns ordered by share of women in the population. Threshold of Rank is 500. We chose this as its mentioned that each election cycle, the top 500 towns on the list are required to reserve the leadership positions for women.

So we observe the towns just above and below the threshold i.e the towns with rank<=500 and rank > 500 to randomly assign.Given the assumption that all the towns ranked <= 500 are assinged treatment a female-headed government and the towns that are ranked > 500 are assigned control i.e a male-headed government.

In Math and words:

We are given that "Each election cycle, the top 500 towns on the list are required to reserve the leadership positions for women (though, be warned – official rules aren't always perfectly followed)". This means that

even if the town in ranked <= 500, the treatment is not assigned and in some cases when the town is ranked > 500 it is not assigned to control. This means that when official rules are not always perfectly followed, we observe a non compliance.

Given that Treatment indicator is D_i and that $D_i \in \{0, 1\}$, $rank_i$ represents rank of town i, and that $D_i = 1$ means treated i.e town i is having a female-headed government and that $D_i = 0$ means town is untreated i.e town i is having a male-headed government.

We get

$Pr(D_i = 1 \mid rank_i <= 500) \; - \; Pr(D_i = 1 \mid rank_i > 500) = k \;\; where \;\; 0 < k < 1$

We can then knowing the non-compliance information implement Fuzzy Regression Discontinuity design, a two stage design to estimate ATE. We chose FRDD becuase we have decided rank variable with a threshold to randomly assign treatment a female-headed government, we doubted existence of possibility of non-compliance and that the probability of treatment varies on both sides of the threshold i.e probability of treatment a female-headed government is higher on one side of threshold rank <= 500 compared to the probability of treatment a female-headed government on the other side of threshold where rank > 500. The implementation is as follows:

Step 1: Estimating the Reduced Form

$Y_i = \alpha + \theta.1[rank_i <= 500] + v_i \; for \; 500 - h <= rank_i <= 500 + h$

Y_i: Total number of public goods h: bandwidth threshold θ: Parameter to estimate the effect of moving from rank > 500 to rank <= 500 on outcome Y_i The Step 1 determines the effect of moving from ran >500 to rank <= 500 on the outcome i.e in our case Public goods.

Step 2: Estimating First stage

$D_i = \alpha + \gamma.1[rank_i <= 500] + v_i \; for \; 500 - h <= rank_i <= 500 + h$

h: bandwidth threshold D_i: Treatment status γ: Parameter that estimates the change in probability of treatment when moving from rank > 500 to rank <= 500
The Step 2 determines the change in probability of treatment when moving from rank >500 to rank <= 500

Step3: Determining τ^{FRD}:

$\hat{\tau}^{FRD} \; = \; \frac{\bar{Y}(500-h<=rank_i<=500) - \bar{Y}(500<=rank_i<=500+h)}{\bar{D}(500-h<=rank_i<=500) - \bar{D}(500<=rank_i<=500+h)} = \; \frac{\hat{\theta}}{\hat{\gamma}}$ h: bandwidth threshold D_i: Treatment status

We can then estimate the τ^{LATE} i.e the Local Average Treatment Effect at threshold 500 for the compliant using the FRD estimator.

$\tau^{LATE} = \hat{\tau}^{FRD}$

Under what conditions would this approach estimate the causal effect of female leadership on public goods provision?

1) For a town i, we assume that rank of the town is the only variable that determines Treatment status D_i that gives $Cov(D_i, rank_i) \neq 0$ thereby $E[D_i \mid rank_i <= 500] \neq E[D_i \mid rank_i > 500]$

2) Assignment is random to a just above and just below the threshold rank 500. This gives

$Y_i(D_i, \; 1[rank_i <= 500]), \;\; D_i(rank_i <= 500), \;\; D_i(rank_i > 500) \;\; \perp \;\; 1[rank_i <= 500]$

3) For a town i, potential outcomes Y_i are equal when the town ranks just below or just above the threshold rand of 500. This is conditional on treatment i.e a female-headed government.

$Y_i(rank_i <= 500, \; D_i) = Y_i(rank_i > 500, \; D_i) \;\; for \;\; D_i \in [0, 1]$ This is also called a Exclusion restriction condition.

4) Monotonicity restriction. We assume that the probability of treatment for a town i ranked <= 500 is greater than the probability of treatment of a town i that is ranked > 500
[$D_i(rank_i <= 500) - D_i(rank_i > 500)$] $>= 0$ $for\ \ any\ \ town\ i$

The monotonicity restictions allows for transition within treated and untreated across the ranks below and above the threshold of 500. With one exception where a treated town i with rank > 500 moves to a untreated status for a rank <= 500

From the above assumptions we get $E[Y_i(1)|rank_i = 500]$ and $E[Y_i(0)|X_i = 500]$ i.e we observe continuity at the threshold rank 500.

As mentioned above, we recover the treatment effect for the towns that transition from a untreated status with rank > 500 to a treated status with rank <= 500. i.e The ones with rank >500 having a male-headed government to a rank <=500 with a female-headed government. In other words , we recover the treatment effects for towns that are compliers. We cant recover the same for non-compliers.

Ans as mentioned above, we can then estimate the τ^{LATE} i.e the Local Average Treatment Effect at rank threshold of 500 for the compliant using the FRD estimator.

$\tau^{LATE} = \hat{\tau}^{FRD}$

Question III F

PROGRAMEVAL likes this idea, and is willing to share data with you to try this out. Use the dataset contained in final_exam_2022.csv. What empirical tests would you like to perform, prior to attempting to estimate the effect of female leadership on public goods provision, to provide evidence in support of the identifying assumption(s)? Perform at least two tests (hint: these should be simple graphical exercises). What do they tell you about the validity of the identifying assumption(s) in this case?

What empirical tests would you like to perform

Would like to perform 4 empirical tests, namely "Outcome across Running variable", "Density of Running variable", "Continuity of Covariates", "Proportion of treatment status across running variable" as shown below:

Outcome across rank: Once we run outcomes across the running variable rank, we will observe if there is any discontinuity at the threshold rank value, in our case 500. Using this analysis we can determine whether RDD is required or not.

Density of rank: To check whether the threshold in itself is actually random or not, we run this test. Once we perfomr density of rank, we can look for the continuity of the density of rank across the threshold rank, in our case 500. We can get to know if the units (towns) are sorted around the running variable (rank) or not.

Covariates continuity or Outcome continuity: We can determine the continuity of potential outcomes by determining the continuity in covariates, in our case per-capita income, number of residents, year of incorporation, average population age etc.. Any difference or discontinuity we observe in the outcomes at the threshold can be said to be based on the shift in treatment status D_i from treated to untreated.

Treatment proportion across rank: By computing the treatment proportion across rank, we can determine the non compliance information. Also helps us to decide whether to implement FRD i.e Fuzzy Regression Discontinuity design or Sharp Regression Discontinuity.

```
data <- read_csv('final_exam_2022.csv')
```

```
## Rows: 5000 Columns: 9
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (9): share_women, list_rank, reservation, female_leader, number_of_resid...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
colnames(data)
```

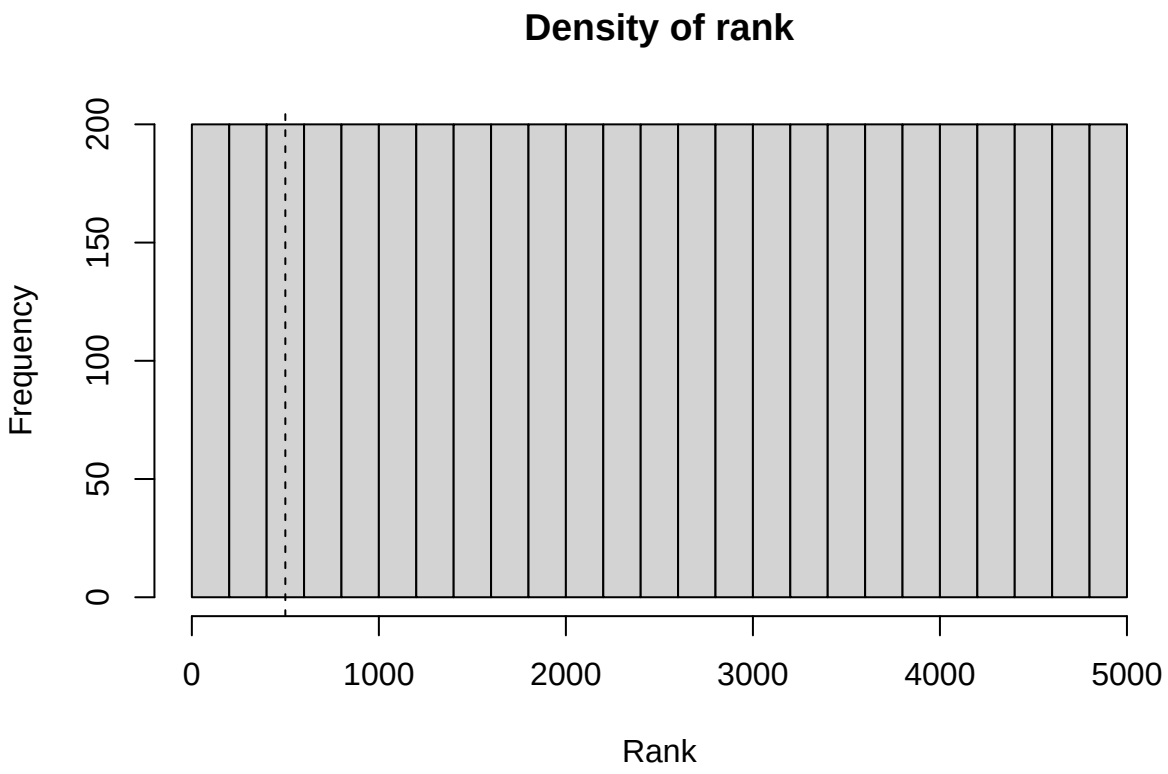
```
## [1] "share_women"                 "list_rank"
## [3] "reservation"                 "female_leader"
## [5] "number_of_residents"         "per_capita_income_rupees"
## [7] "incorp_year"                 "manufacturing_product_share"
## [9] "public_goods_number"
```

```
summary(data)
```

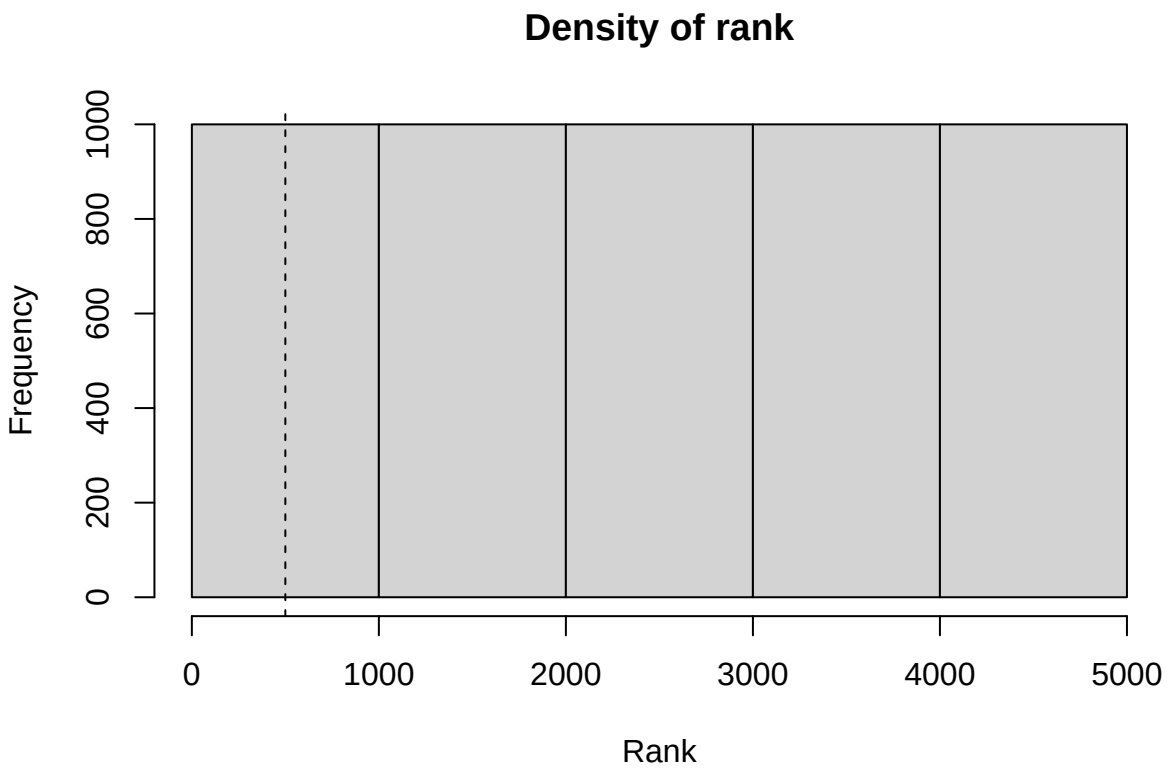
```
##   share_women       list_rank     reservation  female_leader
##  Min.   :0.1521   Min.   :   1   Min.   :0.0   Min.   :0.0000
##  1st Qu.:0.4359   1st Qu.:1251   1st Qu.:0.0   1st Qu.:0.0000
##  Median :0.5008   Median :2500   Median :0.0   Median :0.0000
##  Mean   :0.5014   Mean   :2500   Mean   :0.1   Mean   :0.2676
##  3rd Qu.:0.5700   3rd Qu.:3750   3rd Qu.:0.0   3rd Qu.:1.0000
##  Max.   :0.8844   Max.   :5000   Max.   :1.0   Max.   :1.0000
##  number_of_residents per_capita_income_rupees  incorp_year
##  Min.   :  183       Min.   :    62.9         Min.   :1850
##  1st Qu.:11724       1st Qu.: 49749.2         1st Qu.:1861
##  Median :15140       Median : 88641.2         Median :1873
##  Mean   :15061       Mean   : 78469.7         Mean   :1872
##  3rd Qu.:18428       3rd Qu.:112156.8         3rd Qu.:1884
##  Max.   :32068       Max.   :122756.3         Max.   :1895
##  manufacturing_product_share public_goods_number
##  Min.   :0.0300              Min.   : 0.00
##  1st Qu.:0.3751              1st Qu.: 2.00
##  Median :0.7501              Median : 2.00
##  Mean   :0.6382              Mean   : 4.12
##  3rd Qu.:0.9200              3rd Qu.: 8.00
##  Max.   :0.9200              Max.   :13.00
```

Density of Running Variable i.e Density of Rank:

```
hist(data$list_rank,
     main = "Density of rank",
     xlab = "Rank",
     breaks = 20
     )


abline(v = 500, lty =2)
```



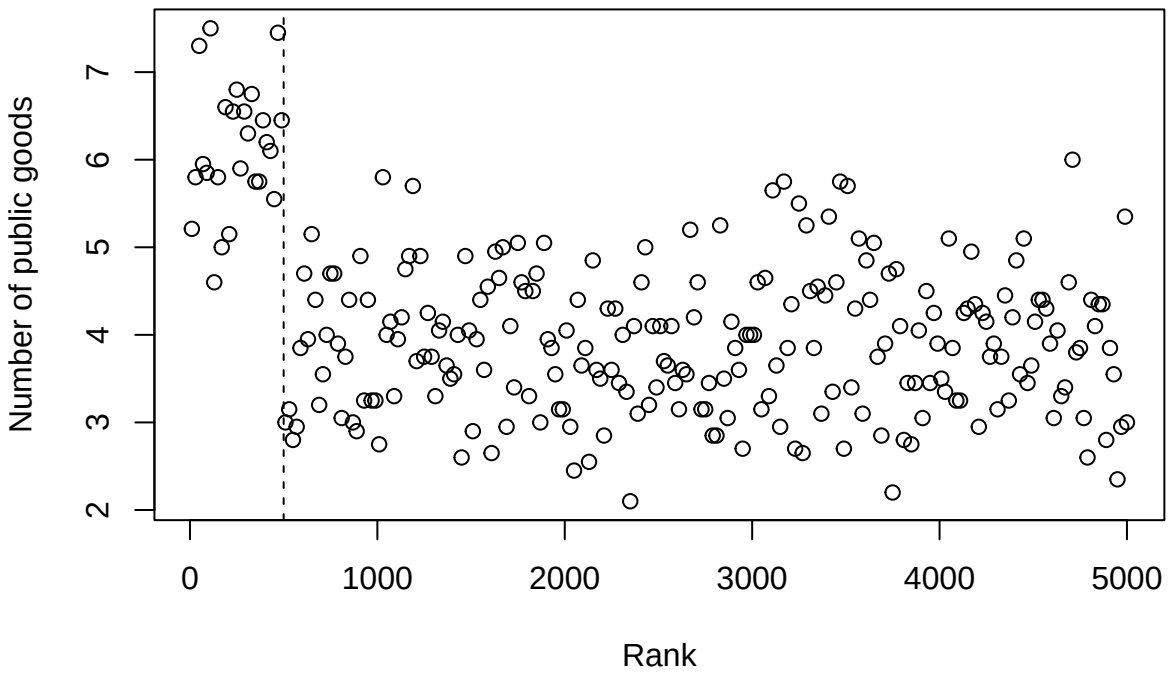

```
hist(data$list_rank,
     main = "Density of rank",
     xlab = "Rank",
     breaks = 5
     )
abline(v = 500,  lty =2)
```



We observe that the density of rank is consistent across and around the threshold rank 500. There are no dips or jumps in the histogram above. Hence we can say that there is rank continuity and that the units (towns) do not differ in order around the threshold value of the running variable (rank).

Test 2: Outcome across Running variable i.e Public Goods across Rank

```
data$bins = floor((data$list_rank - 500)/20)

public_goods = data %>%
               group_by(bins) %>%
               summarise(mean_public_good = mean(public_goods_number,
                                                 na.rm = TRUE),
                         mean_rank = mean(list_rank,
                                          na.rm = TRUE))


plot(public_goods$mean_rank,
     public_goods$mean_public_good,
     xlab="Rank",
     ylab = "Number of public goods")


abline(v=500, lty =2)
```

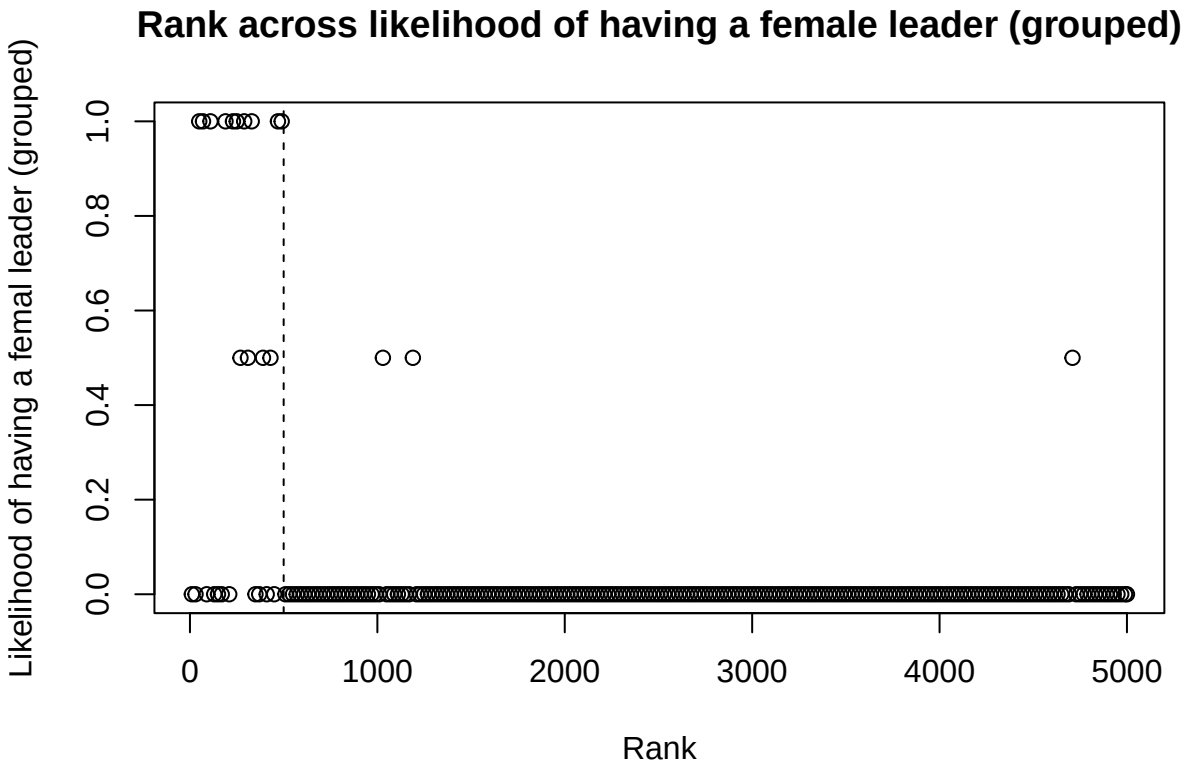


We observe a dip in the public goods when we move from rank <= 500 to a rank >500. Also, we see that overall public goods are lower when rank > 500 whereas they are higher for rank <= 500. From this we can say that public goods (outcome) is more for treated towns i.e towns with female-headed governments.

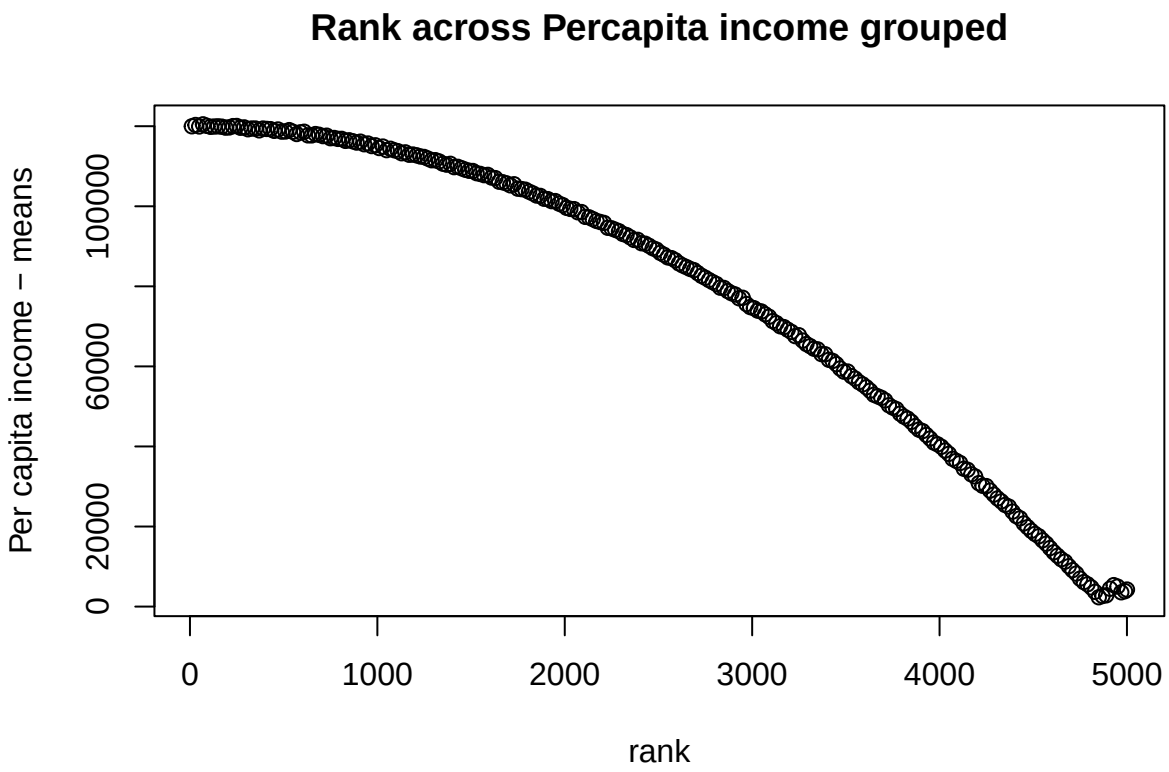
Similarly tests can be run and graphical interpretaions can derived for Covariate smoothness.

Example:

```
covariates_data = data %>%
  group_by(bins) %>%
  summarise(mean_female_leader = median(female_leader, na.rm=TRUE),
            mean_percapita_income = mean(per_capita_income_rupees, na.rm=TRUE),
            mean_public_goods = median(public_goods_number, na.rm=TRUE),
            mean_rank = mean(list_rank, na.rm = TRUE))


plot(covariates_data$mean_rank,
     covariates_data$mean_female_leader,
     main = "Rank across likelihood of having a female leader (grouped)",
     xlab = "Rank",
     ylab = "Likelihood of having a femal leader (grouped)")

abline(v = 500, lty =2)
```




```
plot(covariates_data$mean_rank,
     covariates_data$mean_percapita_income,
     main = "Rank across Percapita income grouped",
     xlab = "rank", ylab = "Per capita income - means")
```



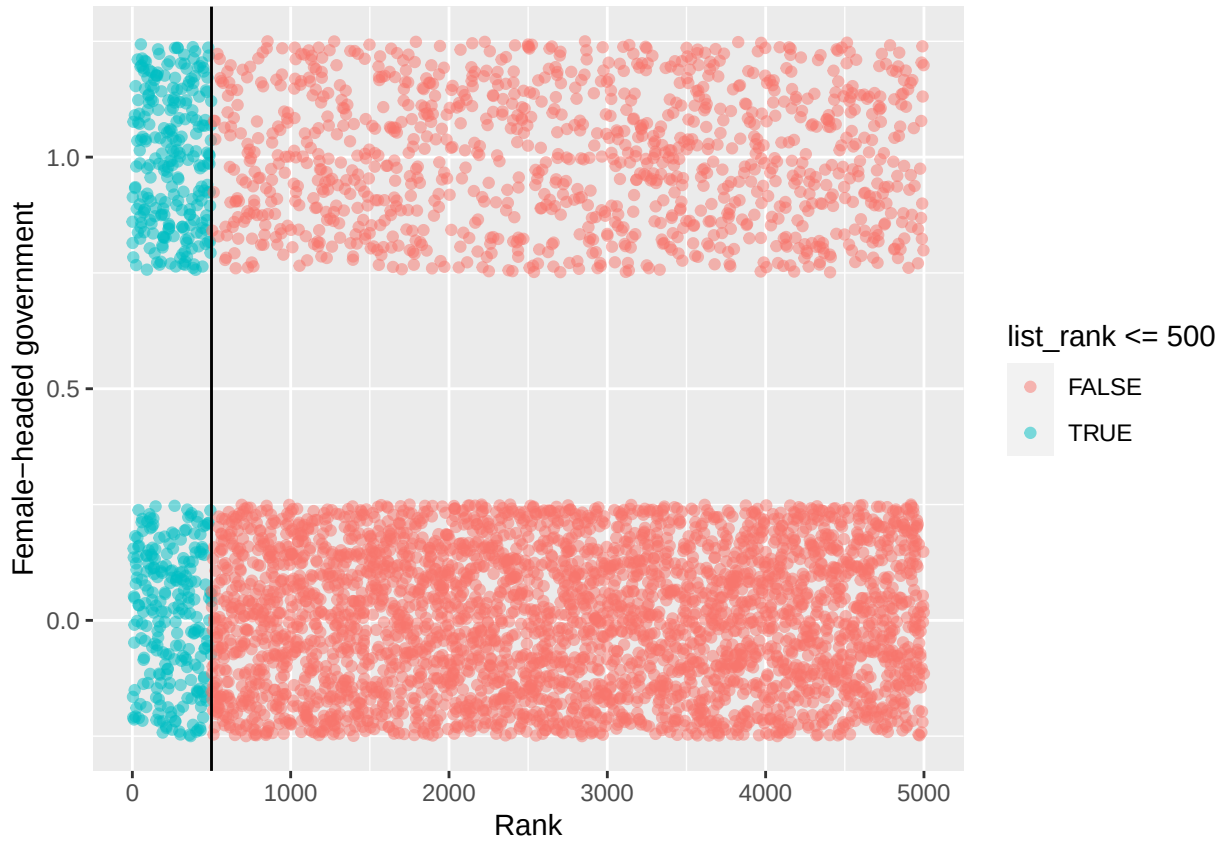
Question III G

Plot the relationship between a town's position on the list and its likelihood of having a female leader. Describe what you're plotting, using a definition from the course. Plot the relationship between the probability of having a female leader and public goods provision. Describe what you're plotting, using a definition from the course. Informed by these plots, write down your preferred regression equation(s) for estimating the causal effect of female leadership on public goods provision. Defend your choice of bandwidth and any functional form choices you make.

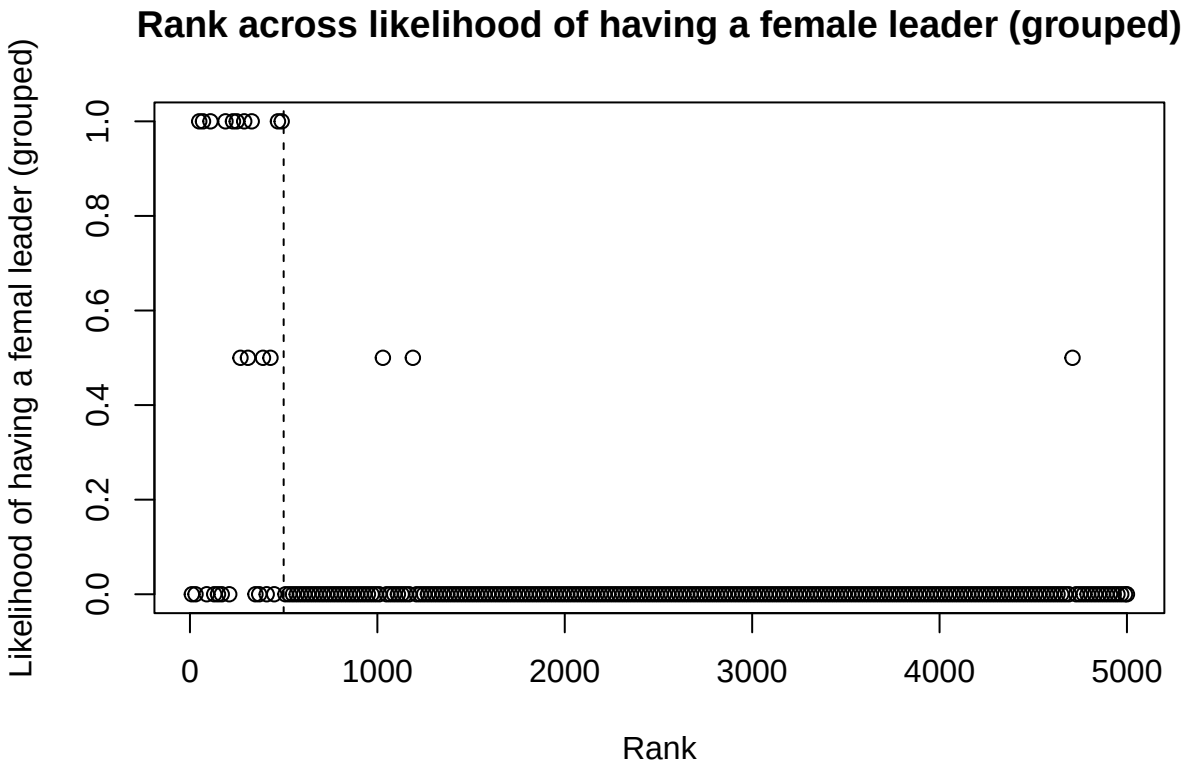
Plot

By plotting the relationship between town's ran and its likelihood of having a female leader, we are determining the Non compliance at threshold of running variable, in our case rank 500 for treated and control groups. In our case, treated is female-headed government and control is no-female-headed government.

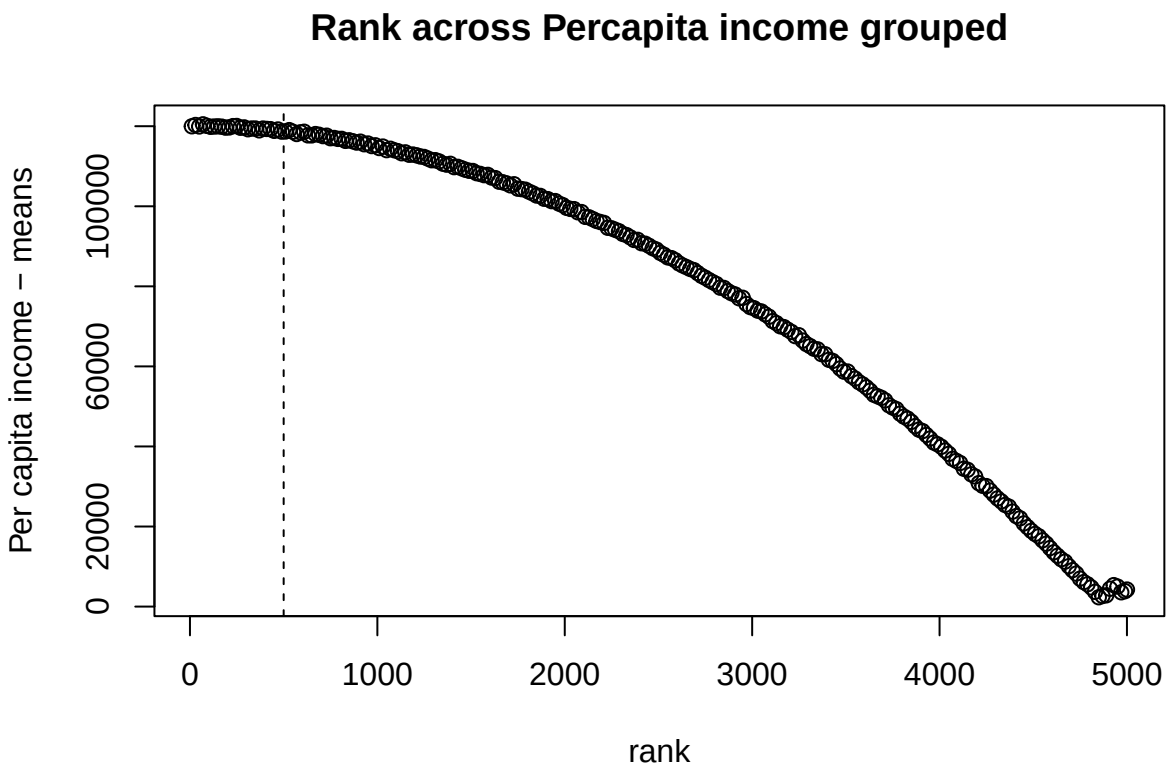
```
ggplot(data, aes(x = list_rank,
                 y = female_leader,
                 color = list_rank <= 500)) +
  geom_point(size = 1.5, alpha = 0.5,
             position = position_jitter(width = 0, height = 0.25, seed = 5555)) +
  labs(x = "Rank", y = "Female-headed government") +
  geom_vline(xintercept = 500)
```




```
covariates_data = data %>%
  group_by(bins) %>%
  summarise(mean_female_leader = median(female_leader, na.rm=TRUE),
            mean_percapita_income = mean(per_capita_income_rupees, na.rm=TRUE),
            mean_public_goods = median(public_goods_number, na.rm=TRUE),
            mean_rank = mean(list_rank, na.rm = TRUE))


plot(covariates_data$mean_rank,
     covariates_data$mean_female_leader,
     main = "Rank across likelihood of having a female leader (grouped)",
     xlab = "Rank",
     ylab = "Likelihood of having a femal leader (grouped)")

abline(v = 500, lty =2)
```




```
plot(covariates_data$mean_rank,
     covariates_data$mean_percapita_income,
     main = "Rank across Percapita income grouped",
     xlab = "rank", ylab = "Per capita income - means")



abline(v = 500, lty =2)
```



We see in the plot that around the rank 500, the treatment status doesnt seem consistent.

We have two variables female_leader(1/0) and rank. A regression of female_leader on rank(<=500 or > 500) i.e female headed government on the reservation status (rank) is the First stage of the Two stage Least squares regression under the Fuzzy Regression discontinuity design research approach.

```
data %>%
  group_by(list_rank < 500, female_leader) %>%
  summarize(count = n()) %>%
  mutate(prop = count / sum(count))
```

```
## `summarise()` has grouped output by 'list_rank < 500'. You can override using the `.groups` argument
```

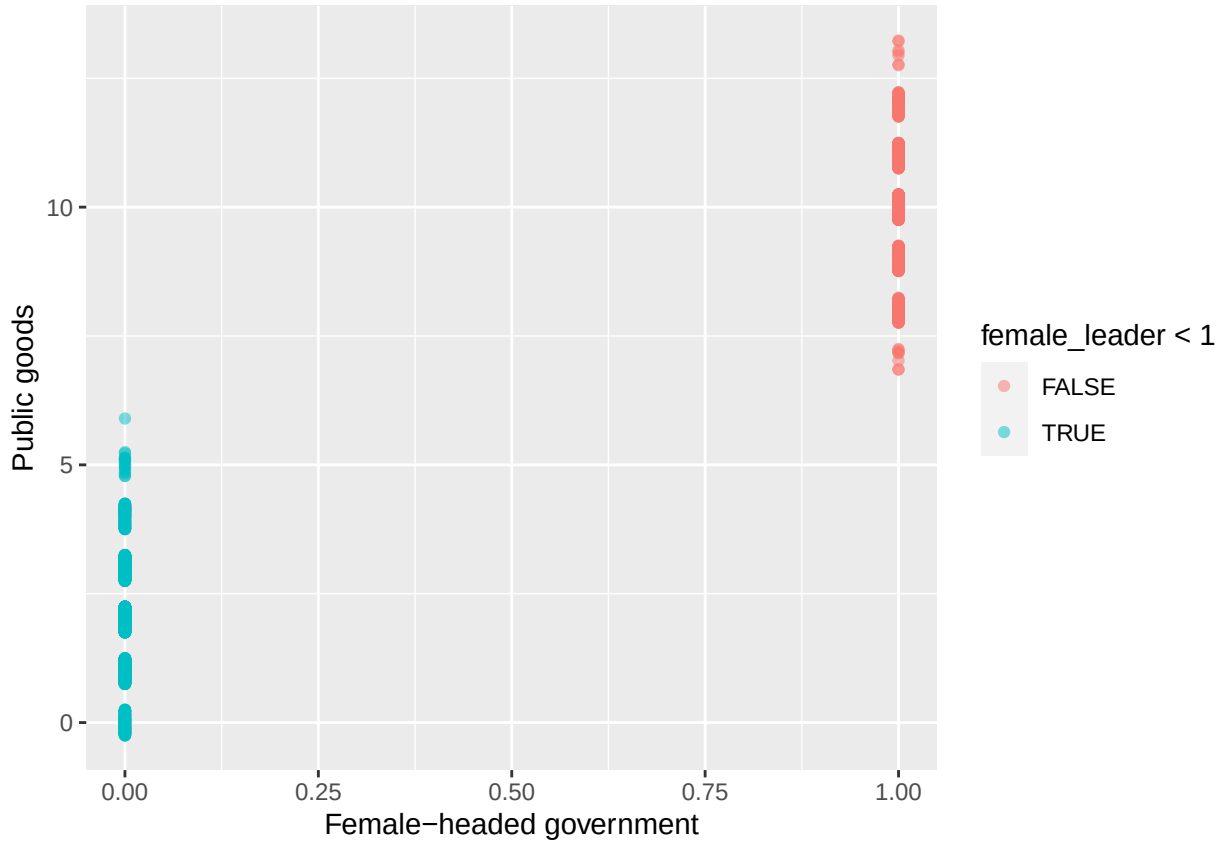
```
## # A tibble: 4 x 4
## # Groups:   list_rank < 500 [2]
##   `list_rank < 500` female_leader count  prop
##   <lgl>                     <dbl> <int> <dbl>
## 1 FALSE                         0  3417 0.759
## 2 FALSE                         1  1084 0.241
## 3 TRUE                          0   245 0.491
## 4 TRUE                          1   254 0.509
```

We observe from the group summary that the non compliance is about 49.09% in the treatment group, similarly 24.08%.

We have two variables female_leader(1/0) and rank. A regression of female_leader on rank(<=500 or > 500) i.e female headed government on the reservation status (rank) is the First stage of the Two stage Least squares regression under the Fuzzy Regression discontinuity design research approach.

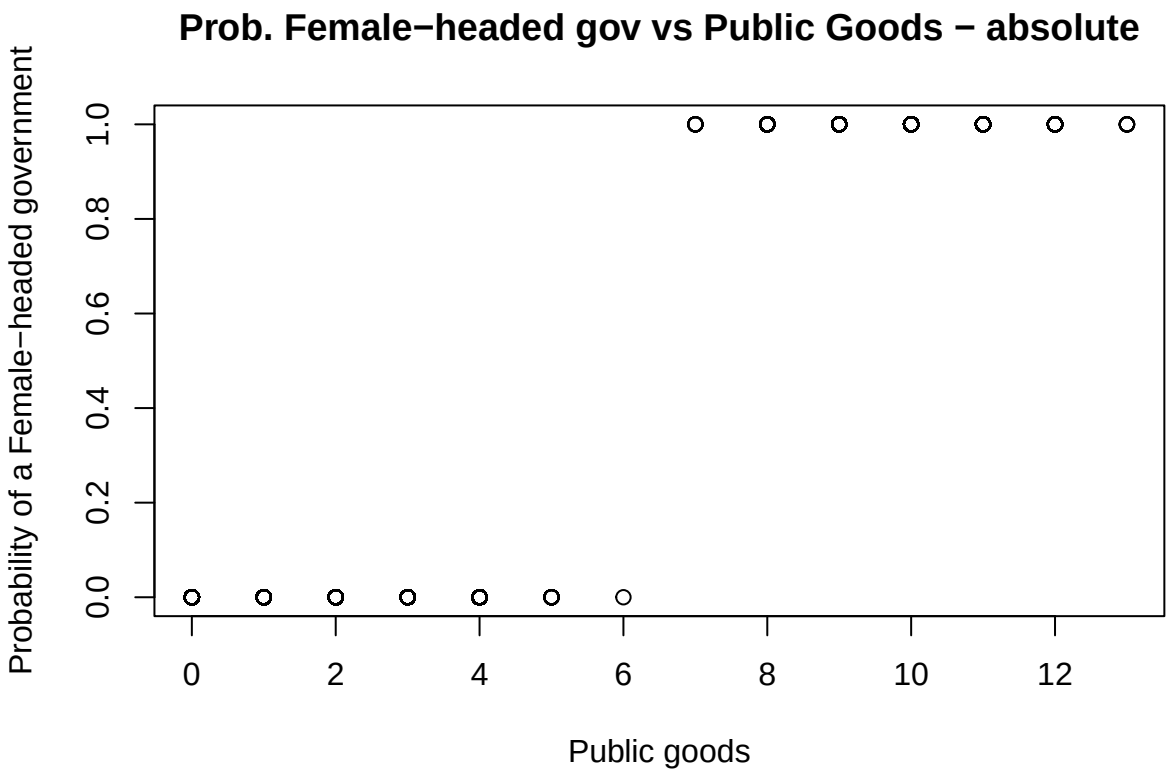
Plot By plotting the relation between probability of having a female leader and public goods provision, we are actually determining the variation of the outcomes Y_i with treatment status D_i. In this context, we are observing the variation of public goods provision with the female-headed government/male-headed government.

```
ggplot(data, aes(x = female_leader, y = public_goods_number , color = female_leader < 1)) +
  geom_point(size = 1.5, alpha = 0.5,
             position = position_jitter(width = 0, height = 0.25, seed = 5555)) +
  labs(x = "Female-headed government", y = "Public goods")
```



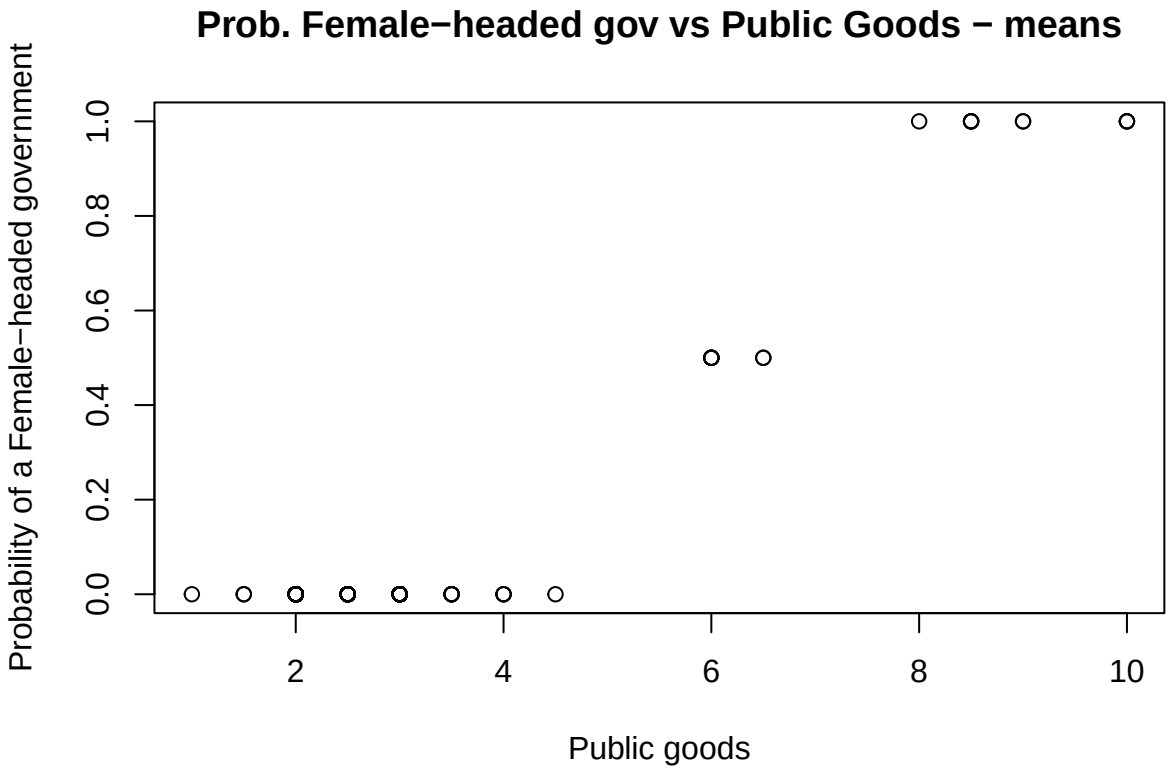
We see that the towns with female-headed governments have more public goods provisions than the ones with no female head.

```
plot(data$public_goods_number, data$female_leader,
     xlab = "Public goods",
     ylab = "Probability of a Female-headed government",
     main = "Prob. Female-headed gov vs Public Goods - absolute")
```



We see that the towns with female-headed governments have more public goods provisions.

```
plot(covariates_data$mean_public_goods, covariates_data$mean_female_leader,
     xlab = "Public goods",
     ylab = "Probability of a Female-headed government",
     main = "Prob. Female-headed gov vs Public Goods - means")
```



We see that the towns with female-headed governments have more public goods provisions.

We have two variables public_goods_number and female_leader(1/0). A regression of public_goods_number on female_leader i.e public goods provision on the female head status is the second stage of Two stage Least squares regression under the Fuzzy Regression discontinuity design research approach.

Regression equations - 2SLS - Fuzzy Regression discontinuity approach:

Estimating First Stage:

$D_i = \alpha + \gamma_1.1[rank_i <= 500] + \gamma_2.rank_i + v_i \ for \ 500 - h <= rank_i <= 500 + h$

h: bandwidth threshold D_i: Treatment status γ_1: Parameter that estimates the change in probability of treatment when moving from rank > 500 to rank <= 500
γ_2: Parameter that estimates the change in probability on treatment status for one unit increase in rank Treatment D_i: Female-headed government / No-female-headed governemt This determines the change in probability of treatment when moving from rank >500 to rank <= 500

Estimating Second Stage:

$Y_i = \alpha + \theta_1.\hat{D}_i + \theta_2.rank_i + v_i \ for \ 500 - h <= rank_i <= 500 + h$

h: bandwidth threshold D_i: Treatment status θ_1: Parameter that estimates the change in outcomes Y_i with treatment status
θ_2: Parameter that estimates the change in our outcomes Y_i with one unit increase in the rank
Treatment D_i: Female-headed government / No-female-headed governemt Y_i: Number of public goods

Justification: The linear functional form mentioned in the stages above is because, the rans that are on the extremes i.e very high or very low are not in intrest. Interest of our analysis lies in the ranks that lie just around the threshold rank value 500. In practice, estimates tend to be very sensitive and moreover the end

points have an outsized impact on the polynomials. Due to these the Fuzzy Regression Continuity model is restricted to a Lineal model provided above.

Bandwidth selection: We can undertake bandwidth selection by using the rdrobust() package in R and also test sensitivity to alternatives.

```
library(rdrobust)
```

```
## Warning: package 'rdrobust' was built under R version 4.1.3
```

```
library(estimatr)
```

```
## Warning: package 'estimatr' was built under R version 4.1.2
```

```
library(modelsummary)
```

```
## Warning: package 'modelsummary' was built under R version 4.1.3
```

```
bw_select = rdbwselect(y = data$public_goods_number,
           x = data$list_rank,
           c = 500,
           all = TRUE)


summary(bw_select)
```

```
## Call: rdbwselect
##
## Number of Obs.                 5000
## BW type                         All
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                  499         4501
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## Unique Obs.                     499         4501
##
## ===================================================
##                   BW est. (h)    BW bias (b)
##             Left of c Right of c  Left of c Right of c
## ===================================================
##      mserd   217.380    217.380    332.681    332.681
##     msetwo   288.639    912.025    367.687   1478.470
##     msesum   237.609    237.609    338.317    338.317
##   msecomb1   217.380    217.380    332.681    332.681
##   msecomb2   237.609    237.609    338.317    338.317
##      cerrd   141.994    141.994    332.681    332.681
##     certwo   188.541    595.742    367.687   1478.470
##     cersum   155.208    155.208    338.317    338.317
##   cercomb1   141.994    141.994    332.681    332.681
##   cercomb2   155.208    155.208    338.317    338.317
## ===================================================
```

In the above results the best bandwidth selected bw est h with Left -217.380 and Right -217.380

Test sensitivity:

```
data_ranked500 <- data %>%
  mutate( ranked_500 = list_rank - 500)

summary(data_ranked500)
```

```
##   share_women       list_rank     reservation  female_leader
##  Min.   :0.1521   Min.   :   1   Min.   :0.0   Min.   :0.0000
##  1st Qu.:0.4359   1st Qu.:1251   1st Qu.:0.0   1st Qu.:0.0000
##  Median :0.5008   Median :2500   Median :0.0   Median :0.0000
##  Mean   :0.5014   Mean   :2500   Mean   :0.1   Mean   :0.2676
##  3rd Qu.:0.5700   3rd Qu.:3750   3rd Qu.:0.0   3rd Qu.:1.0000
##  Max.   :0.8844   Max.   :5000   Max.   :1.0   Max.   :1.0000
##  number_of_residents per_capita_income_rupees  incorp_year
##  Min.   :  183       Min.   :    62.9         Min.   :1850
##  1st Qu.:11724       1st Qu.: 49749.2         1st Qu.:1861
##  Median :15140       Median : 88641.2         Median :1873
##  Mean   :15061       Mean   : 78469.7         Mean   :1872
##  3rd Qu.:18428       3rd Qu.:112156.8         3rd Qu.:1884
##  Max.   :32068       Max.   :122756.3         Max.   :1895
##  manufacturing_product_share public_goods_number      bins
##  Min.   :0.0300              Min.   : 0.00       Min.   :-25.00
##  1st Qu.:0.3751              1st Qu.: 2.00       1st Qu.: 37.00
##  Median :0.7501              Median : 2.00       Median :100.00
##  Mean   :0.6382              Mean   : 4.12       Mean   : 99.55
##  3rd Qu.:0.9200              3rd Qu.: 8.00       3rd Qu.:162.00
##  Max.   :0.9200              Max.   :13.00       Max.   :225.00
##    ranked_500
##  Min.   :-499.0
##  1st Qu.: 750.8
##  Median :2000.5
##  Mean   :2000.5
##  3rd Qu.:3250.2
##  Max.   :4500.0
```

```
fuzzy_1 <- iv_robust(
  public_goods_number ~ ranked_500 + female_leader | ranked_500 + reservation,
  data = filter(data_ranked500, ranked_500 >= -217 & ranked_500 <= 217)
)
tidy(fuzzy_1)
```

```
##           term     estimate    std.error statistic      p.value     conf.low
## 1   (Intercept) 1.8967537578 0.1982103304  9.569399 8.380790e-20  1.507177199
## 2    ranked_500 0.0001315568 0.0006354295  0.207036 8.360793e-01 -0.001117361
## 3 female_leader 8.5281026456 0.5213563843 16.357530 3.733889e-47  7.503392048
##     conf.high  df             outcome
## 1 2.286330317 432 public_goods_number
## 2 0.001380475 432 public_goods_number
## 3 9.552813244 432 public_goods_number
```

```
fuzzy_2 <- iv_robust(
  public_goods_number ~ ranked_500 + female_leader | ranked_500 + reservation,
  data = filter(data_ranked500, ranked_500 >= -110 & ranked_500 <= 110)
)
tidy(fuzzy_2)
```

```
##            term     estimate   std.error statistic      p.value     conf.low
## 1   (Intercept)  2.069051354 0.190712413 10.849065 3.303432e-22  1.693175192
## 2    ranked_500 -0.001965006 0.001611032 -1.219719 2.238894e-01 -0.005140197
## 3 female_leader  7.863195343 0.502639569 15.643805 1.737707e-37  6.872540201
##     conf.high  df             outcome
## 1 2.444927516 218 public_goods_number
## 2 0.001210185 218 public_goods_number
## 3 8.853850484 218 public_goods_number
```

```
fuzzy_3 <- iv_robust(
  public_goods_number ~ ranked_500 + female_leader | ranked_500 + reservation,
  data = filter(data_ranked500, ranked_500 >= -220 & ranked_500 <= 220)
)
tidy(fuzzy_3)
```

```
##            term     estimate    std.error  statistic      p.value     conf.low
## 1   (Intercept) 1.8677345968 0.2100929490  8.8900394 1.606457e-17  1.454818992
## 2    ranked_500 0.0003220399 0.0006779439  0.4750244 6.350064e-01 -0.001010388
## 3 female_leader 8.6145565177 0.5538507871 15.5539303 9.494690e-44  7.526021021
##     conf.high  df             outcome
## 1 2.280650201 438 public_goods_number
## 2 0.001654467 438 public_goods_number
## 3 9.703092014 438 public_goods_number
```

```
fuzzy_4 <- iv_robust(
  public_goods_number ~ ranked_500 + female_leader | ranked_500 + reservation,
  data = filter(data_ranked500, ranked_500 >= -330 & ranked_500 <= 330)
)
tidy(fuzzy_4)
```

```
##            term     estimate    std.error statistic      p.value      conf.low
## 1   (Intercept) 1.7855550691 0.1840060265  9.703786 6.745113e-21  1.4242452907
## 2    ranked_500 0.0003476204 0.0003662724  0.949076 3.429304e-01 -0.0003715833
## 3 female_leader 8.6335265033 0.4679326180 18.450363 1.405698e-61  7.7147053441
##     conf.high  df             outcome
## 1 2.146864848 658 public_goods_number
## 2 0.001066824 658 public_goods_number
## 3 9.552347662 658 public_goods_number
```

```
modelsummary(list("Fuzzy bandwidht = +/-110" = fuzzy_2,
                  "Fuzzy bandwidht = +/-217" = fuzzy_1,
                  "Fuzzy bandwidht = +/-220" = fuzzy_3,
                  "Fuzzy bandwidht = +/-330" = fuzzy_4
                  ),
             stars = TRUE)
```

From the results obtained in the sensitivity test , we choose a bandwidth h of +/- 217.

	Fuzzy bandwidht = +/-110	Fuzzy bandwidht = +/-217	Fuzzy bandwidht = +/-220	Fuzzy ba
(Intercept)	2.069***	1.897***	1.868***	
	(0.191)	(0.198)	(0.210)	
ranked_500	−0.002	0.000	0.000	
	(0.002)	(0.001)	(0.001)	
female_leader	7.863***	8.528***	8.615***	
	(0.503)	(0.521)	(0.554)	
Num.Obs.	221	435	441	
R2	0.931	0.934	0.933	
R2 Adj.	0.930	0.933	0.932	
Std.Errors	HC2	HC2	HC2	
statistic.weakinst				
p.value.weakinst				
statistic.endogeneity				
p.value.endogeneity				
statistic.overid				
p.value.overid				

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Question III H

Finally, estimate the causal effect of female leadership on public goods provision. What do you find? Interpret your results. Advise PROGRAMEVAL: should they expand female leadership to all towns?

Causal effect of female leadership (i.e rank <= 500) on public goods provision

Coefficients summary at bandwidth levels of +/- 217

```
tidy(fuzzy_1)
```

```
##           term     estimate    std.error statistic      p.value     conf.low
## 1   (Intercept) 1.8967537578 0.1982103304  9.569399 8.380790e-20  1.507177199
## 2    ranked_500 0.0001315568 0.0006354295  0.207036 8.360793e-01 -0.001117361
## 3 female_leader 8.5281026456 0.5213563843 16.357530 3.733889e-47  7.503392048
##     conf.high  df             outcome
## 1 2.286330317 432 public_goods_number
## 2 0.001380475 432 public_goods_number
## 3 9.552813244 432 public_goods_number
```

Coefficient of female_leader = 8.528. Interpretation of this coefficient, means that a female-headed government in the town i will increase an average of 8.528 in Public Goods provision for the compliers within the bandwidth of 217.

Hence we get $\hat{\tau}^{FRD} = 8.528$

We now run the First stage 2SLS and determine the strength and significance of the IV Rank within the bandwidths.

```
reg1 <- lm( female_leader ~ reservation+
              ranked_500 + number_of_residents +
              per_capita_income_rupees +
```

```
              incorp_year + manufacturing_product_share,
              data = filter(data_ranked500,
                            ranked_500 >= -217 &
                            ranked_500 <= 217)
)


summary(reg1)
```

```
##
## Call:
## lm(formula = female_leader ~ reservation + ranked_500 + number_of_residents +
##     per_capita_income_rupees + incorp_year + manufacturing_product_share,
##     data = filter(data_ranked500, ranked_500 >= -217 & ranked_500 <=
##         217))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5984 -0.3237 -0.1873  0.4764  0.8959
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -4.680e+00  4.209e+00  -1.112   0.2668
## reservation                  3.694e-01  8.793e-02   4.202 3.23e-05 ***
## ranked_500                  -4.862e-05  1.455e-03  -0.033   0.9734
## number_of_residents          7.835e-06  4.729e-06   1.657   0.0982 .
## per_capita_income_rupees     1.486e-05  2.077e-05   0.716   0.4746
## incorp_year                  1.462e-03  1.681e-03   0.870   0.3848
## manufacturing_product_share  1.575e+00  4.763e+00   0.331   0.7411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4577 on 428 degrees of freedom
## Multiple R-squared:  0.1066, Adjusted R-squared:  0.09403
## F-statistic: 8.508 on 6 and 428 DF,  p-value: 9.616e-09
```

We get coefficient value for reservations 0.349, number_of_residents 7.981e-06,per_capita_income_rupees 1.313e-05, manufacturing_product_share 8.985e-01 with and F statistic of 9.059 on 6 and 434 DF. We observe a p-value of 2.41e-09

The F-statistic value is < 10 indicating that the IV used rank is weak in determining the female_leader. This means that rank is not the only variable that determines the treatment status. Thus the first condition that the running variable is the only variable that determiness the treatment status doesnt satisfy anymore.

Thus $\hat{\tau}^{FRD}$ and $\hat{\tau}^{LATE}$ are not equal.

We can run the second stage to further analyse this.

```
reg2 <- lm( public_goods_number ~ female_leader +
              ranked_500 + number_of_residents +
              per_capita_income_rupees +
              incorp_year + manufacturing_product_share,

            data = filter(data_ranked500,
```

```
                          ranked_500 >= -217 &
                          ranked_500 <= 217)
)
summary(reg2)
```

```
##
## Call:
## lm(formula = public_goods_number ~ female_leader + ranked_500 +
##     number_of_residents + per_capita_income_rupees + incorp_year +
##     manufacturing_product_share, data = filter(data_ranked500,
##     ranked_500 >= -217 & ranked_500 <= 217))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3952 -0.9176 -0.0491  0.8331  3.2567
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1.364e+01  9.432e+00  -1.446   0.1489
## female_leader                8.159e+00  1.061e-01  76.881   <2e-16 ***
## ranked_500                  -5.228e-03  3.212e-03  -1.628   0.1043
## number_of_residents         -2.361e-05  1.062e-05  -2.223   0.0268 *
## per_capita_income_rupees     4.132e-05  4.655e-05   0.888   0.3752
## incorp_year                  4.555e-03  3.769e-03   1.209   0.2275
## manufacturing_product_share  1.721e+01  1.066e+01   1.614   0.1072
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.025 on 428 degrees of freedom
## Multiple R-squared:  0.937,  Adjusted R-squared:  0.9361
## F-statistic:  1061 on 6 and 428 DF,  p-value: < 2.2e-16
```

We used rank as IV.

Coefficient for female_leader = 8.159 , number_of_residents -2.361e-05, per_capita_income_rupees 4.132e-05 at the p-value of 2.2e-16. More over we get a F statistic of 1061 on 6 and 428 DF.

The coefficient for female_leader means that, means that a female-headed government in the town i will increase an average of 8.159 in Public Goods provision for the compliers within the bandwidth of 217.

As in first stage analysis we found that the Rank variable is a weak IV in determining the treatment, we cannot say that $\hat{\tau}^{FRD}$ and $\hat{\tau}^{LATE}$ are equal. We also saw the difference in numbers obtained from the senstivity test and the coefficient in the 2nd stage Fuzzy regression discontinuity design.

As we cannot with clarity say that the female leadership in towns would result in increasing the number of Public goods and thus the provision of the public goods, I DONOT recommend PROGRAMEVAL to expand female leadership to all towns.

Bonus

Find an example of a popular press article describing a study using causal language, when, given what you've learned in this quarter, this is likely not appropriate. Use a few sentences to describe the study and the main problem of the study through the lens of this course. Attach the article in PDF form to your exam submission.