R Final IV and Regression Discontinuity

Tytus Wilam

12/13/2019

(1) Hypothetical real-life scenario. We have a randomized encouragement design combined with a sharp eligibility cutoff. Imagine a study where 1000 high school seniors are given are encouraged do go to college by a tuition subsidy z. Some of the students go to college, others don't. Only students from households with below 50.000 yearly income are eligible for the subsidy. Students from the other (not-encouraged) group are also allowed to go to college but receive no subsidy. Some students undergo treatment d, which is going to college right after high school. In the student population there are never takers who would not go to college regardless of the subsidy, compliers who undergo treatment only after being encouraged, and always takers who would have gone to college regardless of the subsidy. The proportion of compliers and always takers increases with income. We measure outcome for all students which is a binary indicator of graduating from college within 10 years (y). As researchers, we are interested in the causal effect of the subsidy on students who would not have attempted higher education without it (CACE, complier average causal effect).

(2) DGP. For World A and C I simulate income as a mixture of normal distributions which are truncated at 0 and 250 to avoid negative incomes and outliers. The distributions are the following:

$$N(20, 50)$$

$$N(230, 15)$$

In World B I add a third normal distribution that is truncated at the eligibility cutoff to violate the assumption that the forcing variable and the cutoff are independent.

In all worlds I get eligibility based on the income. Only students from households with less than 50k income are eligible. In all worlds I simulate compliance status as a binomial distribution with probabilities rising with each try. The higher the income, the likelier someone is to be a complier or always taker, with always takers concentrating among the richest.

I simulate binary potential outcomes also using a binomial distribution. The probability of successful outcome increases with treatment and income. In world A and B the probability p of successful outcome is:

$$p_{ij} = 0.5d_{ij} + 0.002 * income_i$$

Where i indicates the person, j indicates eligibility status, and d indicates treatment. Treatment increases the probability of success by 0.5.

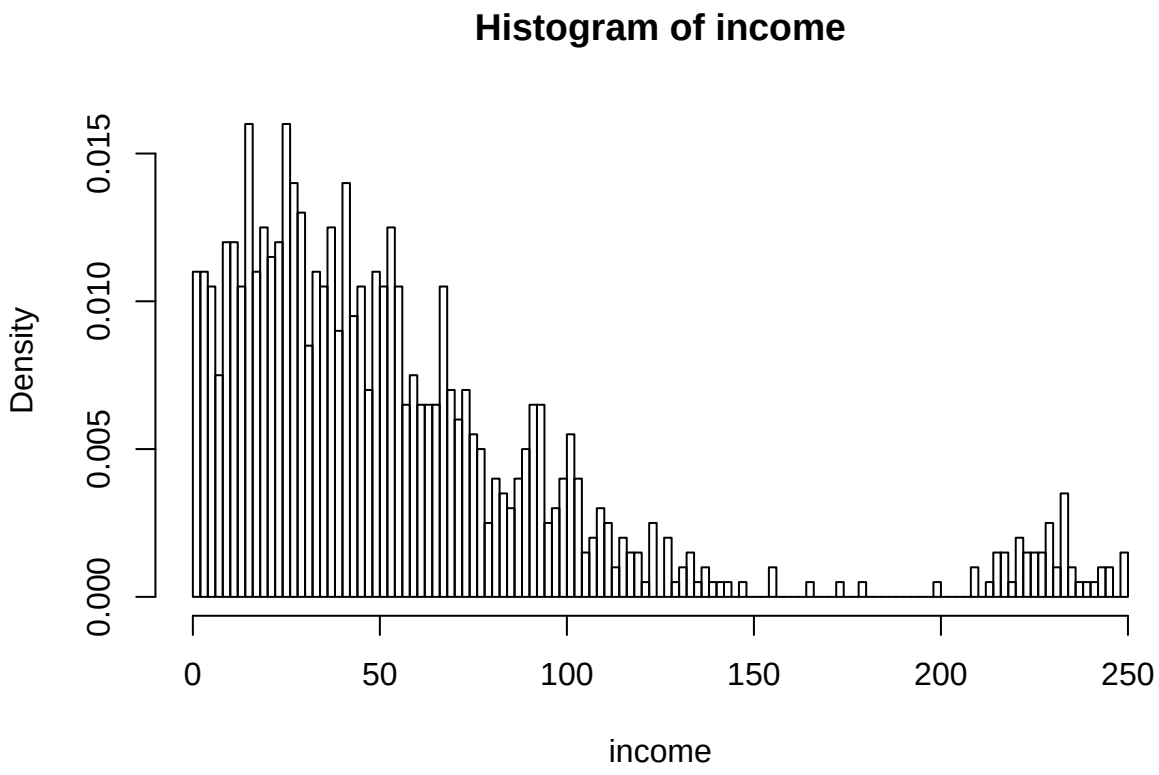
In World C eligibility independently contributes to a successful outcome. For those who are eligible for treatment the probability of a successful outcome is:

$$p_i = 0.2d_i + 0.3 * z + 0.002 * income_i$$

```
### (3)(i) DGP code for World A

## I simulate the forcing variable: income
set.seed(1234)
income <- c(rtruncnorm(950, a=0, b=200, 20, 50), rtruncnorm(50, a=0, b=250, 230, 15))
## I determine the eligibility based on median income of 50k
eligible <- ifelse(income < 50, 1, 0)
## I simulate compliance status; values -- {0: never taker, 1: complier, 2: always taker}
c <- rbinom(1000, 2, seq(from=0.25, to=0.9, length.out = 1000))
## I simulate potential outcomes for the treatment variable
```

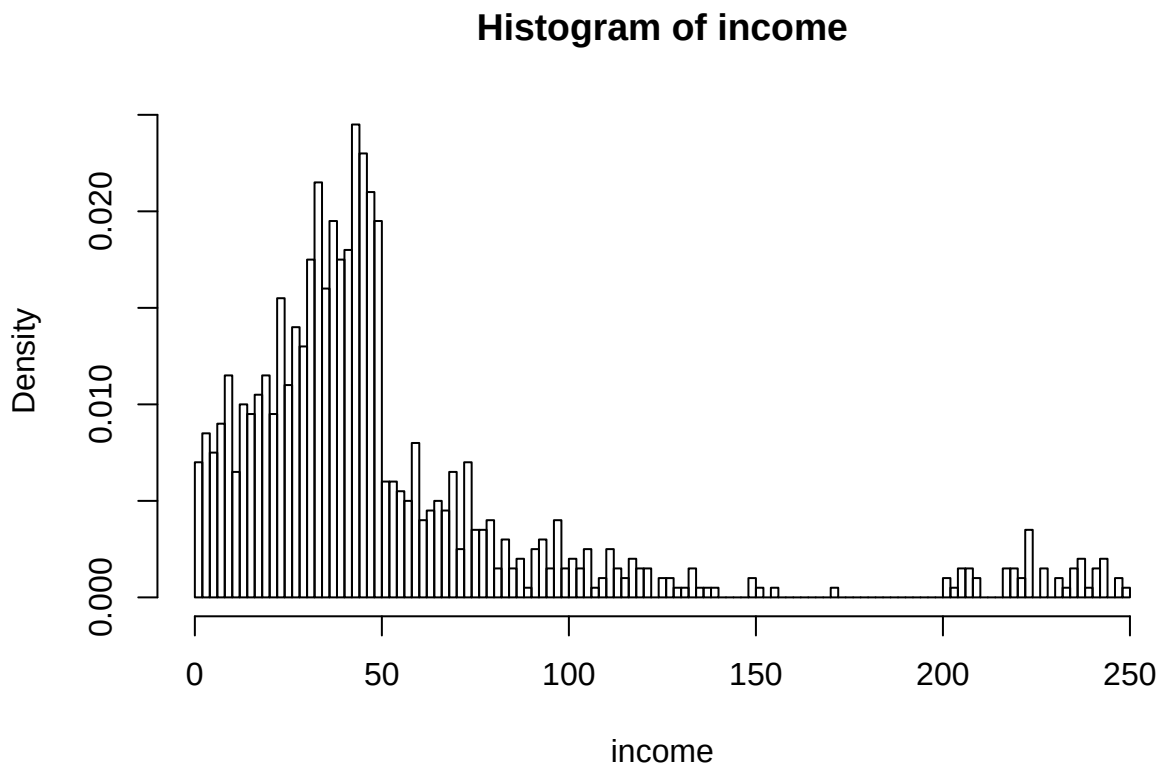
```r
d0 <- ifelse(c==0 | c==1, 0, 1)
d1 <- ifelse(c==0, 0, 1)
## I simulate potential outcomes for a single outcome variable
# I first correlate income with outcome
prob.y0 <- d0*0.5 + income*0.002
prob.y1 <- d1*0.5 + income*0.002
# Then I simulate potential outcomes
y0 <- rbinom(1000, 1, prob.y0)
y1 <- rbinom(1000, 1, prob.y1)
## Binary encouragement variable. Note that it's the same as eligibility.
z <- ifelse(eligible == 1, 1, 0)
## Binary treatment receipt variable
d <- ifelse(z==1, d1, d0)
## Observed results
y <- ifelse(d==1, y1, y0)
data.full <- tbl_df(data.frame(income, eligible, c, d0, d1, y0, y1, z, d, y))
data.obs <- select(data.full, income, eligible, z,d,y)
hist(income, freq=FALSE, breaks=100)
```




```r
### (3)(ii) DGP code for World B where the assumption about
# independence of the forcing varaible and the cutoff score is violated.

## I simulate the forcing variable: income
income <- c(rtruncnorm(600, a=0, b=200, 20, 50),rtruncnorm(350, a=0, b=50, 50, 20), rtruncnorm(50, a=0,
## I determine the eligibility based on median income of 50k
```

```r
eligible <- ifelse(income < 50, 1, 0)
## I simulate compliance status; values -- {0: never taker, 1: complier, 2: always taker}
c <- rbinom(1000, 2, seq(from=0.25, to=0.9, length.out = 1000))
## I simulate potential outcomes for the treatment variable
d0 <- ifelse(c==0 | c==1, 0, 1)
d1 <- ifelse(c==0, 0, 1)
## I simulate potential outcomes for a single outcome variable
# I first correlate income with outcome
prob.y0 <- d0*0.5 + income*0.002
prob.y1 <- d1*0.5 + income*0.002
# Then I simulate potential outcomes
y0 <- rbinom(1000, 1, prob.y0)
y1 <- rbinom(1000, 1, prob.y1)
## Binary encouragement variable. Note that it's the same as eligibility.
z <- ifelse(eligible == 1, 1, 0)
## Binary treatment receipt variable
d <- ifelse(z==1, d1, d0)
## Observed results
y <- ifelse(d==1, y1, y0)
dataB.full <- tbl_df(data.frame(income, eligible, c, d0, d1, y0, y1, z, d, y))
dataB.obs <- select(dataB.full, income, eligible, z,d,y)
hist(income, freq=FALSE, breaks=100)
```




```r
### (3)(iii) DGP for World C where exclusion restriction is violated
# and eligibility has an effect on outcome independent of treatment.
```

```
income <- c(rtruncnorm(950, a=0, b=200, 20, 50), rtruncnorm(50, a=0, b=250, 230, 15))
## I determine the eligibility based on median income of 50k
eligible <- ifelse(income < 50, 1, 0)
## I simulate compliance status; values -- {0: never taker, 1: complier, 2: always taker}
c <- rbinom(1000, 2, seq(from=0.25, to=0.9, length.out = 1000))
## I simulate potential outcomes for the treatment variable
d0 <- ifelse(c==0 | c==1, 0, 1)
d1 <- ifelse(c==0, 0, 1)
## I simulate potential outcomes for a single outcome variable
# I first correlate income with outcome
prob.y0 <- d0*0.5 + income*0.002
prob.y1 <- d1*0.2 + eligible*0.3 + income*0.002
# Then I simulate potential outcomes
y0 <- rbinom(1000, 1, prob.y0)
y1 <- rbinom(1000, 1, prob.y1)
## Binary encouragement variable. Note that it's the same as eligibility.
z <- ifelse(eligible == 1, 1, 0)
## Binary treatment receipt variable
d <- ifelse(z==1, d1, d0)
## Observed results
y <- ifelse(d==1, y1, y0)
dataC.full <- tbl_df(data.frame(income, eligible, c, d0, d1, y0, y1, z, d, y))
dataC.obs <- select(dataC.full, income, eligible, z,d,y)
```

(4) (a) Description of method and the estimand I use regression discontinuity and IV. Regression discontinuity can be conceptualized as a randomized experiment with noncompliance. IV design can be conceptualized as a randomized experiment. I combine RD and IV the two approaches. The estimand is the complier average treatment effect (CACE) at the cutoff.

I use two stage least squares to estimate the percentage of compliers and then to estimate CACE. Instead of randomly assigning treatment, I use the fact that eligibiblity is arbitrarily assigned based on the forcing varaible. For this reason, I am only able to estimate CACE near the cutoff.

```
### (4)(b)
### Code to estimate CACE near the cutoff using 2SLS
### with 30k bandwith
# for World A

data.obs.A.bw30 <- filter(data.obs, income >20 & income < 80)
fit <- lm(d ~ z, data.obs.A.bw30)
d_predict <- fit$fitted.values
fit <- lm(y ~ d_predict, data.obs.A.bw30)
CACE.A.2SLS.30 <- fit$coefficients[2]

# for World B
data.obs.B.bw30 <- filter(dataB.obs, income >20 & income < 80)
fit <- lm(d ~ z, data.obs.B.bw30)
d_predict <- fit$fitted.values
fit <- lm(y ~ d_predict, data.obs.B.bw30)
CACE.B.2SLS.30 <- fit$coefficients[2]

# for World C
data.obs.C.bw30 <- filter(dataC.obs, income >20 & income < 80)
fit <- lm(d ~ z, data.obs.C.bw30)
```

```
d_predict <- fit$fitted.values
fit <- lm(y ~ d_predict, data.obs.C.bw30)
CACE.C.2SLS.30 <- fit$coefficients[2]

### with 3k bandwith
# for World A

data.obs.A.bw <- filter(data.obs, income >47 & income < 53)
fit <- lm(d ~ z, data.obs.A.bw)
d_predict <- fit$fitted.values
fit <- lm(y ~ d_predict, data.obs.A.bw)
CACE.A.2SLS <- fit$coefficients[2]

# for World B
data.obs.B.bw <- filter(dataB.obs, income >47 & income < 53)
fit <- lm(d ~ z, data.obs.B.bw)
d_predict <- fit$fitted.values
fit <- lm(y ~ d_predict, data.obs.B.bw)
CACE.B.2SLS <- fit$coefficients[2]

# for World C
data.obs.C.bw <- filter(dataC.obs, income >47 & income < 53)
fit <- lm(d ~ z, data.obs.C.bw)
d_predict <- fit$fitted.values
fit <- lm(y ~ d_predict, data.obs.C.bw)
CACE.C.2SLS <- fit$coefficients[2]

### Code to estimate CACE near the cutoff using ivreg and 30k bandwith

fit <- ivreg(y ~ d, ~z, data.obs.A.bw30)
CACE.ivreg.error.A <- fit$sigma
CACE.ivreg.A <- fit$coefficient[2]

fit <- ivreg(y ~ d, ~z, data.obs.B.bw30)
CACE.ivreg.error.B <- fit$sigma
CACE.ivreg.B <- fit$coefficient[2]

fit <- ivreg(y ~ d, ~z, data.obs.C.bw30)
CACE.ivreg.error.C <- fit$sigma
CACE.ivreg.C <- fit$coefficient[2]
```

(5) Assumptions I have separate assumptions for IV and for RD. 5 assumptions for IV: ## Assumption 1: SUTVA – Stable Unit Treatment Value Assumption (IV and RD assumption) Potential outcomes of each person are unaffected by the treatment status and outcomes of other persons. This assumption is likely violated in the proposed setting because the study might affect the composition of the student body which could affect the likelihood of completing college for individuals.

Assumption 2: Random Assignment of the Instrument (IV assumption)

Instrument is randomly assigned. (i.e. ignrability of the instrument). This assumption holds here only close to the cutoff and only if assumpiton 3 is also satisfied.

Assumpiton 3: Independence of forcing variable and cutoff (RD assumption)

The cutoff score and the forcing variable are determined independently of one another. This appears to be satisfied in this study. We would expect this to be violated if people knew about the program and underreported their incomes to qualify for it, or if there were other programs that had a similar cutoff.

Assumpiton 4: Exclusion Restriction (IV assumption)

Instrument affects the outcome only through the treatment, i.e. if instrument were different but outcome stayed the same the outcome would also stay the same. I violate this assumption by proposing that the researcher operationalized treatment as going to college in the first year after high school, which failed to capture the causal effect of the subsidy instrument on going to college in later years.

Assumption 5: Monotonicity (IV assumption)

We expect no defiers. This should be satisfied although defiers remain possible in principle.

Assumption 6: Nonzero average causal effect of instrument on treatment (IV assumption)

This assumption seems plausible in this setting.

Assumption 7: Average outcome is a continuous function of the forcing variable (RD assumption)

This is satisfied in this setting.

(6) Attractive display of results

```
###(6) table of results

data.frame(WorldA=c(CACE.A.2SLS.30, CACE.A.2SLS, CACE.ivreg.A, CACE.ivreg.error.A), WorldB=c(CACE.B.2SLS
```

```
##                               WorldA    WorldB    WorldC
## CACE with 30k bandwith     0.4636058 0.3903961 0.6860196
## CACE with 6k bandwith      0.5395683 0.5732218 0.7093023
## CACE with ivreg and 30k    0.4636058 0.3903961 0.6860196
## ivreg CACE standard error 0.4333611 0.4427040 0.4463394
```

Discussion of the two-stage least squares complier average causal effect for world A with 30k bandwith: treatment d causally increases the likelihood of outcome y by 46% for compliers at the eligibility cutoff. In the real world scenario, this means that those students who would not have gone to college if they did not receive the encouragement z and who are right below the eligibility cutoff of 50k, have a 46% likelihood of completing college if they go to college. The entire effect can be causally attributed to the treatment, because those studnets would not go to college in the counterfactual scenario. This effect is causal, i.e. going to college causes those students to complete college.

(7) Assumptions and bias.

In World C I violated the Instrumental Variable design exclusion restriction assumption. The instrument (z) affects the outcome (y) independently of its effect on the treatment (d). The real world scenario could

be that we measure treatment as going to college right after high school and that we measure as outcome college completion within 10 years. If eligibility for a tuition subsidy influences students to take up college more than one year after completing high school, then the instrument affects the outcome in a way not accounted by our operationalization of treatment. Because of the violation of the assumption, IV produces biased results in all three methods.

In World B I violated the Regression Discontinuity assumption that the eligibility cutoff is determined independently of the forcing variable. This could happen if the encouragement is causing people to underreport their incomes or earn less so that they qualify for the subsidy. It is likely that a large scale government program similar to the study here described would generate a distortion of this kind. The assumption violation introduces some bias. Violating this assumption yields an estimand that is and less efficient. What is interesting, is that a wider bandwith halps to make the estimand more accurate.

In Worlds A and B the estimates are less biased when I use a wider bandwidth. This is because the smaller bandwidth has too few observations.

(8) Lessons learned from simulation I have learned: (a) It is possible to fruitfully combine IV and RD design if (b) Exclusion restriction is a fundamental assumption for IV and if we violate it our results become meaningless. (c) That an operationalization of the treatment that is too narrow can lead to a violation of the exclusion restriction. (d) In some situations where the underlying model is linear, the bias from violating of the RD assumption that the cutoff is determined independently of the forcing variable can be mitigated by increasing the bandwith.

I have also learned some lessons from engaging with the assignment. I had to think harder about the DGP process that would be compatible with both IV and RD approaches which gave me a deeper understanding of both.