Microeconometrics II

Homework I

Professor: André Portela   Student: Rafael F. Bressan

2020-11-06

Part 1: Instrumental variables

You have been provided with a sample of first-born children aged between 10 and 18 years old living with both of their parents. The sample was constructed from Census 2010 . Each entry in the dataset corresponds to a first-born child from both spouses. There can be multiple first-born children in a given family/couple if the first birth from both spouses was a multiple birth. A detailed data dictionary can be found in `dicionario.xlsx`. Your goal is to estimate the causal impact of the number of children in the family (variable `family_number_children`) on the years of education of the couple's first-born child(ren) (variable `first_child_years_of_education`)

Question 1 In this item, we will explore the identifying power of the monotone treatment response (MTR) and monotone treatment selection (MTS) assumptions.

(a) State the MTR and MTS assumptions - ***in the most plausible direction*** *- in this context. Does economic theory have any predictions on the validity of these assumptions? Hint: See the discussion in Ponczek and Souza (2012) .*

Both MTR and MTS definitions are made in Manski and Pepper (2000) , they respectively are:

MTR: Let T be an ordered set. For each $j \in J$,

$w_2 \geq w_1 \implies y_j(w_2) \geq y_j(w_1)$.

MTS: Let T be an ordered set. For each $w \in W$,

$u_2 \geq u_1 \implies E[y(w)|z = u_2] \geq E[y(w)|z = u_1]$.

where W denotes the treatment status, y_j is the potential outcome for person $j \in J$.

In this context, the MTR assumption is most plausible when the number of children in the family **reduces** years of education of the first born child. That is:

$$\text{MTR: } w_2 \geq w_1 \implies y_j(w_2) \leq y_j(w_1) \tag{1}$$

considering that w_i is the number of children in the family and the outcome y_j is years of education of the first born.

In a similar way, the MTS assumption in this context will state that potential outcomes for first born children in bigger families are **lower** than the ones in smaller families.

$$\text{MTS: } w_2 \geq w_1 \implies E[y(w)|W = w_2] \leq E[y(w)|W = w_1] \tag{2}$$

Economic theory, as presented in Ponczek and Souza (2012) Introduction section, predicts that larger families, with a higher number of children will have to share its scarce resources among all children in household, thus leaving the first born with a lower level of education when compared to smaller families. Although, the empirical evidence is mixed, many of the literature for developed economies show no effect of family size on child quality while for developing countries, the negative hypothesized effect is found.

(b) Report a table with average years of schooling by number of children, as well as the frequency of each value of variable "number of children" in the sample. You may want to use individual sample weights (variable person_weight) in estimation in order to better account for the population of interest.

The total population represented by this sample is 8,274,695, considering a person's sample weight.

Table 1: Years of schooling and number of children by family size.

Children	Years Educ.	Effective Obs.
1	5.85	2281111.58
2	5.92	3547212.08
3	5.66	1587638.98
4	5.11	520484.33
5	4.65	199463.71
6	4.31	83861.10
7	4.17	34215.65
8	3.92	13763.26
9	3.75	4533.67
10	3.56	1624.90
11	4.55	525.70
12	4.73	150.14
13	3.39	80.26
14	4.20	17.29
15	3.27	7.38
16	3.00	4.96

(c) Compute the upper and lower bounds on the ATE of increasing the number of children in the family from 1 to 2,2 to 3,3 to 4,4 to 5,5 to 6 etc. Compute 95% confidence intervals to these bounds using the bootstrap. You should draw samples of households with replacement in order to properly account for the sampling process. You may want to set the probability of sampling a household proportional to the household weight (variable `household_weight`) in order to better replicate the sampling process.

First we will derive some results **for binary** treatment, following the MTR and MTS specifications from equations (1) and (2) to get an intuition of what changes from the usual MTR and MTS assumptions. When we impose the MTR assumption in (1), we get a zero **upper bound** on ATE, while the MTS assumption yields a **lower bound** that is equal to the naive difference of means between the two groups. We see that upper and lower bound were interchanged in this new specification. Putting the two assumptions together we have that:

$$E[y_j|W = \bar{w}_n] - E[y_j|W = w_n] \leq \text{ATE} \leq 0 \tag{3}$$

where the treatment $\bar{w}_n$ refers to a family with **more** than n children, and w_n is a family with n children or less, such that $P(W = \bar{w}_n) = \pi$ and $P(W = w_n) = 1 - \pi$.

We begin with the proof that MTR assumption implies the zero upper bound.

Proposition 1. *Given the MTR assumption on* (1) *the upper bound – UB – on the average treatment effect – ATE – is zero.*

Proof. Suppose the treatment is binary with two levels, $w_2 \geq w_1$. The probability of being assigned to w_2 is π. Then the MTR assumption in (1) implies that $y_j(w_2) \leq y_j(w_1)$, and we have the following two inequalities,

$E[y_j(w_2)|w_1] \leq E[y_j|w_1]$ and $E[y_j|w_2] \leq E[y_j(w_1)|w_2]$

The ATE has the following observational-counterfactual decomposition,

$E[y_j(w_2) - y_j(w1)] = \pi E[y_j|w_2] - (1-\pi)E[y_j|w_1] - \pi E[y_j(w_1)|w_2] + (1-\pi)E[y_j(w_2)|w_1]$

Making use of the inequalities in the observational-counterfactual decomposition to obtain an upper bound for the ATE we have that:

$$\begin{aligned} \text{ATE} = E[y_j(w_2) - y_j(w1)] &\leq \pi E[y_j|w_2] - (1-\pi)E[y_j|w_1] - \pi E[y_j|w_2] + (1-\pi)E[y_j|w_1] \\ &= 0 \end{aligned}$$

□

Now we prove that the MTS assumption implies a lower bound on the ATE equal to $E[y_j|w_2] - E[y_j|w_1]$.

Proposition 2. *Given the MTS assumption on* (2) *the lower bound – LB – on the average treatment effect is equal to the difference of means between the treatment groups, ATE* $\geq E[y_j|w_2] - E[y_j|w_1]$.

Proof. Suppose once again the treatment is binary with two levels, $w_2 \geq w_1$. The probability of being assigned to w_2 is π. Then the MTS assumption in (2) implies that potential outcomes for treatment group at $W = w_2$ are lower than in group with $W = w_1$, and we have the two inequalities,

$E[y_j(w_1)|w_2] \leq E[y_j|w_1]$ and $E[y_j|w_2] \leq E[y_j(w_2)|w_1]$

by the observational-counterfactual decomposition the ATE has a lower bound given by:

$$\begin{aligned} \text{ATE} = E[y_j(w_2) - y_j(w1)] &\geq \pi E[y_j|w_2] - (1-\pi)E[y_j|w_1] - \pi E[y_j|w_1] + (1-\pi)E[y_j|w_2] \\ &= E[y_j|w_2] - E[y_j|w_1] \end{aligned}$$

□

Therefore, for a multi-level treatment we can expect the same phenomenon to occur, the upper and lower bound computations will be flipped. This in turns defines that our **upper bound** on ATE will be zero, while the lower bound can be computed by eq. 9.19 from Manski (2009).

$$\Delta(s,t) \leq \sum_{t'>t} E(y|w=t')P(w=t') + E(y|w=t)P(w \leq t) - \sum_{s'<s} E(y|w=s')P(w=s') - E(y|w=s)P(w \geq s) \tag{4}$$

Below we present a table with empirical mean of years of study and the distribution number of children across families, a la Table I in Manski and Pepper (2000).

Table 2: Mean of years of study by number of children

w	E[y|w]	P(w)	Size
1	5.8532	0.2757	2281111.5757
2	5.9240	0.4287	3547212.0783
3	5.6572	0.1919	1587638.9786
4	5.1053	0.0629	520484.3335
5	4.6514	0.0241	199463.7111
6	4.3053	0.0101	83861.0976
7	4.1697	0.0041	34215.6500
8	3.9180	0.0017	13763.2559
9	3.7451	0.0005	4533.6739
10	3.5628	0.0002	1624.9029
11	4.5510	0.0001	525.7000
12	4.7339	0.0000	150.1365
13	3.3924	0.0000	80.2619
14	4.1965	0.0000	17.2925
15	3.2710	0.0000	7.3811
16	3.0000	0.0000	4.9640

In table 3 we have the estimated lower bound for the causal effect and the 2.5% and 97.5% bootstrap quantiles, as in Manski and Pepper (2000) Table II.

Table 3: Lower bound on years of study for first child.

		Lower Bound on Delta(s,t)		
s	t	Estimate	2.5% quant.	97.5% quant.
1	2	-0.0913	-0.0651	-0.0323
2	3	-0.3306	-0.3086	-0.2732
3	4	-0.7463	-0.7305	-0.6721
4	5	-1.1245	-1.1277	-1.0341
5	6	-1.4462	-1.5007	-1.3424

As a complementary analysis of our resampling method, we present below the histograms of bootstrapped lower bounds by treatment level t.

Question 2 In this item, we will use an instrumental-variable approach in estimating the causal impact of the number of children on years of schooling of the first child. We will follow the approach in Ponczek and Souza (2012), whereby we first restrict our sample to families with two or more births from the couple (variable `family_number_births` $\geq$ 2). We then propose to instrument the number of children with `second_birth_ismultiplebirth`, which indicates whether the second birth of the couple was a multiple birth [3]

(a) Make the sample restrictions previously discussed. How many second births are multiple births?

When considering only the families with more than one birth we end up with 644453 observations and the number of second births that are multiple is **9446**.

(a) Under which assumptions does an instrumental variable approach identify a treatment effect in our setting? What treatment effect? Do these assumptions seem plausible to you? Why? Looking at the dataset, do you think any controls should be included? Why?

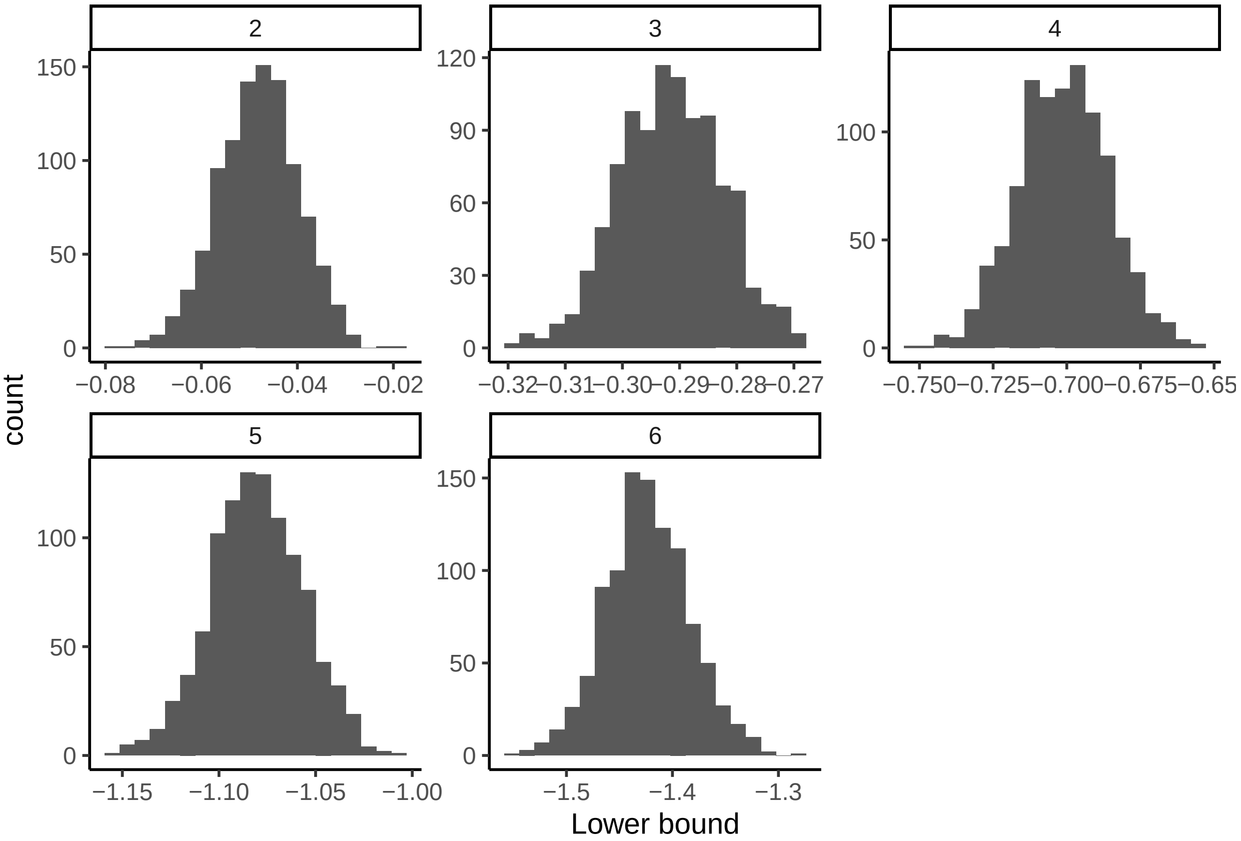

Figure 1: Distribution of bootstrapped lower bounds

First we need the usual IV assumptions of relevance and exclusion. Then, for a causal interpretation of the estimated parameter we need that the instrument must be as good as randomly assigned after controlling for relevant variables. Suppose we have an instrument Z for treatment W, with $Z_i, W_i \in \{0,1\}$ then:

i) Relevance: $E[Z_i W_i] \neq 0$

ii) Exclusion: $E[Z_i \varepsilon_i] = 0$

iii) As good as randomly assigned: $Y_i(1), Y_i(0) \perp Z_i | \mathbf{X_i}$

In the present case, the instrument variable of choice is a multiple birth for the second born child. Relevance is satisfied since a multiple birth is clearly related to the family size. The exclusion restriction of multiple births has been debated in the literature, but many recognized authors have chosen this instrument on the assumption that having twins is largely random, thus, not related to any possible omitted variable that is related to the outcome years of education of first born child. Finally, from the previous discussion about the exclusion restriction, we do believe this instrument is as good as randomly assigned.

Only if the instrument Z_i were **perfectly** correlated to the the treatment W_i we would be able to identify the ATE. Since this is not usually the case, the instrument is only moderately correlated (i.e. relevant) to the treatment, ATE is not identifiable from this setting. This is the case of imperfect compliance, and in order to identify some other effect from the data we need a fourth assumption:

iv) Monotonicity: $W_i(1) \geq W_i(0)$

That is, the effect of the instrument on the treatment points toward the same direction for every unit. If we assume monotonicity, then the Local Average Treatment Effect – LATE – is identified. The LATE has the meaning of the average treatment effect on the subpopulation of **compliers** only. The compliers are those individuals who would take the treatment if induced to do so (by the instrument variable), but otherwise would refrain.

Actually, the twins instrument is special in the sense it does not allow a subpopulation called **never-takers**. If a multiple second birth is present, there is no way this family ends up with only two children, it has to be three or more. In this special case, LATE coincides with the average treatment effect on the untreated – ATU, see Angrist and Pischke (2008), section 4.4.2 for a discussion.

Therefore, for the instrumental variable `second_birth_ismultiplebirth` we do believe all assumptions are plausible and the LATE can be estimated from data.

The set of control variables could potentially be empty, since our instrument is as good as randomly assigned. If the researcher chooses to control for some characteristics, it is more related to improve the estimate precision than satisfying assumption iii) above.

We do not necessarily think any control **should** be include, although, if we can control for other characteristics that helps in determine the outcome, this inclusion would improve our estimate by reducing the estimator's variance. Such variables could be, age of the first child, whether the first child is a girl. The indicative of first birth is multiple is also a good candidate for control variable, since it may explain the schooling of the multiple first born.

(b) Estimate the treatment effect using 2SLS. Include any covariates you regarded as necessary in the previous item. Cluster your standard errors at the household level (you may also want to weight observations by the person weight). Is the instrument relevant? Why? Comment on your results.

We estimate two models. The first one is a single regression model of years of education on number of children, instrumented by multiple second birth. The second model adds the covariates child age, girl and multiple first birth.

	Model 1	Model 2
(Intercept)	6.662	-3.505
	(0.099)	(0.074)
family_number_children	-0.349	-0.397
	(0.037)	(0.032)
first_child_age		0.714
		(0.002)
first_child_is_girl		0.615
		(0.007)
first_birth_ismultiplebirth		-0.209
		(0.049)

Note: Cluster robust standard errors reported in parentheses. Cluster variable is 'id_household'.

In order to assess the instrument's relevance we report the first stage **F statistic**. For model 1 and model 2 this statistic is, respectively, 5917 and 6143, showing that the chosen instrument is relevant.

The inclusion of control variables didn't change significantly the parameter estimate, as expected. Since our instrument is valid, relevant and randomly assigned, controlling for other relevant variables just improves precision, but do not alter the point estimate. We notice however, that all three control variables are significant, but robust standard error for model 2 improved only slightly.

(c) Conduct a test for weak instruments. Are your instruments weak? In what sense? Hint: See Section 4 in Andrews, Stock, and Sun (2019) .

We conduct the Olea and Pflueger (2013) test for weak instrument. The endogenous variable is `family_number_children` and the instrument is `second_birth_ismultiplebirth`, while we add three control variables, `first_child_age`, `first_child_is_girl`, `first_birth_ismultiplebirth`. The reported effective F-statistic was 5498.1354258. Since the authors suggested rule of thumb for a 5% test that the worst-case relative bias of 2SLS exceeds 10% is 23.1 for their corrected F-statistic, we **do reject** the null hypothesis that `second_birth_ismultiplebirth` is a weak instrument for `family_number_children`.

This test uses the first definition of Stock and Yogo of a weak instrument, that is, a instrument is said to be weak when the worst-case bias of two-stage least squares exceeds 10% of the worst-case bias of OLS.

(d) Report Anderson-Rubin confidence intervals. How do they compare to (b)?

We run the AR confidence interval for model 2, with control variables. The confidence interval found is [-0.4, -0.4], while this interval for the same model in item b) was [-0.4443, -0.3498], very close to each other.

(e) Compare your results with the estimates found in Question 1.

We can see that the confidence interval found using the Anderson-Rubin method is tighter than the bounds found in item 1 c) using Manski's approach, in general. Although, for smaller number of children in the family, the lower bound in table 3 are tighter than the confidence interval just found.

Part 2: Regression discontinuity design

For this part of the list, you have been provided with data from Amarante et al. (2016), who studies the effect of a cash-transfer program in Uruguay on health outcomes at birth. According to the authors, "the Uruguayan Plan de Atención Nacional a la Emergencia Social (PANES) was a temporary social assistance program targeted to the poorest 10 percent of households in the country, implemented between April 2005 and December 2007 ." Eligibility was defined via a baseline survey conducted with applicants. A probit model for the likelihood of falling below a critical per capita income level was estimated using baseline data, and households whose predicted probability exceeded some threshold were eligible to the program. However, due to imperfect enforcement of the rules of the program, some noneligible mothers did actually receive the cash transfer, whereas some eligible mothers failed to do so.

You have been provided with a dataset where each entry corresponds to a pair (birth,mother) during the program duration. The treatment indicator variable is treat. The eligibility dummy is `eligible` $= \mathbf{1}\{\text{running} > 0\}$, where running is the predicted probability of falling to poverty, already subtracted of the threshold for program eligibility. File `dic_amarante.pdf` contains the description of additional variables in the dataset.

1. State the assumptions required for the identification of a treatment effect using the discontinuity described above. What do these assumptions mean in this context? What is the interpretation of the treatment effect identified under these assumptions?

The discontinuity just described is a fuzzy one, the probability of receiving the cash transfer even when the running variable is lower than the specified threshold is above zero while that probability in below one if the individual's running variable is over the threshold.

Therefore, the key assumptions for the Fuzzy RDD model are:

i) Potential outcomes are continuous in the running variable x at the cutoff value $\bar{x}$, such that the following limits exist

$$\lim_{x \uparrow \bar{x}} E[Y_i | x = \bar{x}] \text{ and } \lim_{x \downarrow \bar{x}} E[Y_i | x = \bar{x}]$$

this means that we can compare unities just below the cutoff which had not received treatment and unities above the cutoff which did have been treated, in other words, we can compare the outcomes of compliers.

ii) Potential treatments must be discontinuous at $x = \bar{x}$. This ensures there is still a discontinuity on the probability of getting treated at $x = \bar{x}$, although not sharp. Also, this hypothesis is related to the identification of causal effect, since it makes the so called first stage different from zero.

$$\lim_{x' \uparrow \bar{x}} P(D_i = 1 | x = x') \neq \lim_{x' \downarrow \bar{x}} P(D_i = 1 | x = x')$$

This hypothesis is the monotonicity hypothesis without being explicit about the direction of the relation between the running variable and treatment.

iii) Independence of potential outcomes from treatment status at $x = \bar{x}$. This is the usual hypothesis of conditional unconfoundedness, but this time we need it to hold at the cutoff value.

$$Y_i(1), Y_i(0) \perp D_i | x = \bar{x}$$

Since the Fuzzy RDD is similar in nature to an IV approach, we also need two assumptions from instrumental variables, relevance of the running variable and the exclusion restriction, where the running variable does not directly affects the outcome.

Without hypothesis iii) of conditional independence, the treatment effect estimated under a Fuzzy RDD is the Local Average Treatment Effect – LATE – which is the effect on compliers, for those with $x = \bar{x}$. If we impose conditional independence of potential outcomes from treatment at the cutoff value, then the Average Treatment Effect – ATE – is recovered[1].

2. Report a discontinuity plot between the running variable (x-axis) and program participation (y-axis), as well as a discontinuity plot between the running variable (x-axis) and low birthweight (variable `bajo2500`). What do these plots tell you?

[1] From a preliminary draft of A Practical Introduction to Regression Discontinuity Designs: Extensions, p. 88.

Since we have so many observed units, the scatter plots asked would be clogged with points at $Y = \{0, 1\}$ and would be uninformative. Thus, we opted instead, to cut the running variable into bins and compute y-axis variable's average on a given bin. The point estimate and the ± one standard error are provided in the following plots.

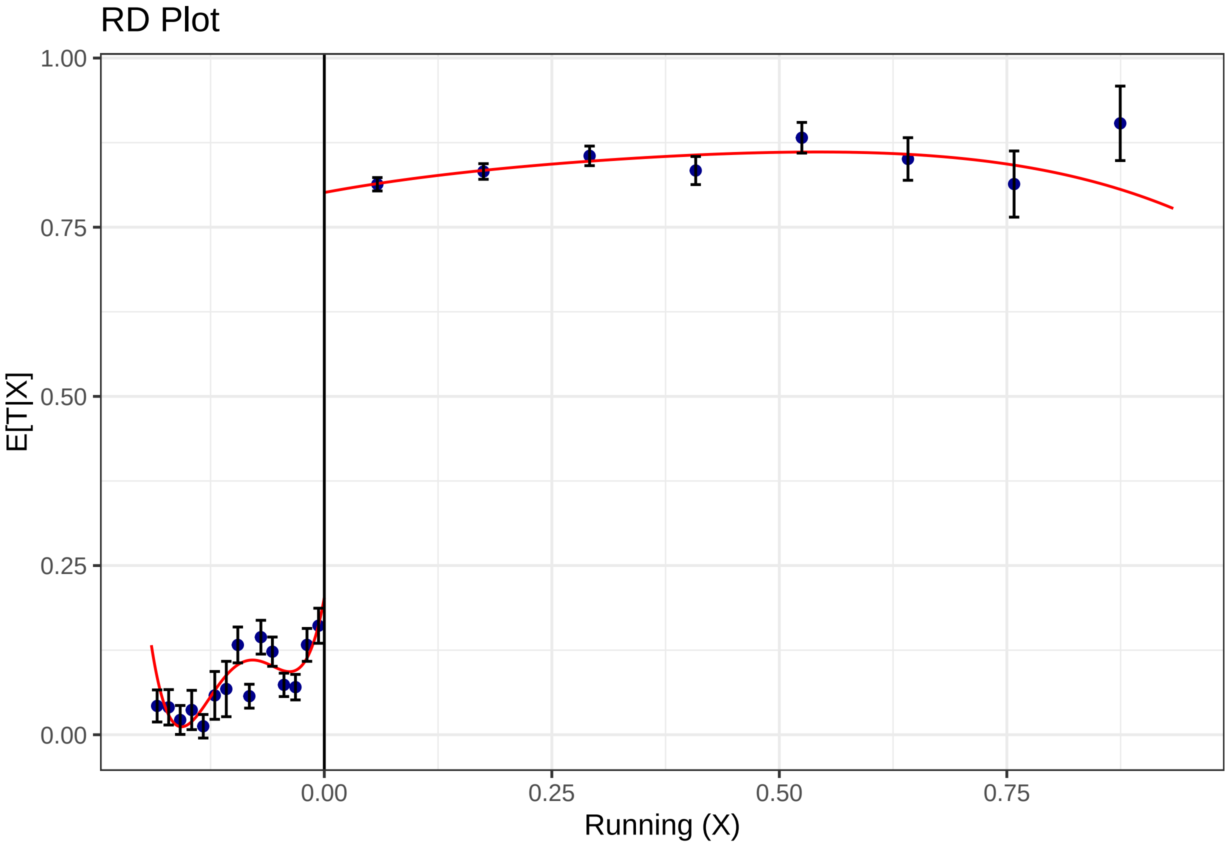

Figure 2: Probability of getting treated.

Figure 2 provides a picture for the discontinuity of the probability of getting treated at $x = 0$, although it is not sharp, there is a clear jump on the point estimate at the cuttof value. On the other hand, the probability of having a low weight child does not appear to change at all at the specified cutoff value, as can be seen in Figure 3.

3. Estimate the effect of program participation on low birthweight by local linear regression. Precisely state the bandwidth selection method used, the choice of kernel, as well as whether bias correction was employed. What is the first-stage relation, at the cutoff, between program eligibility and program participation? Is it statistically significant? Comment on your second stage results.

Below we present the results (second stage) of our estimations using two kernels, triangular and uniform, for comparison.

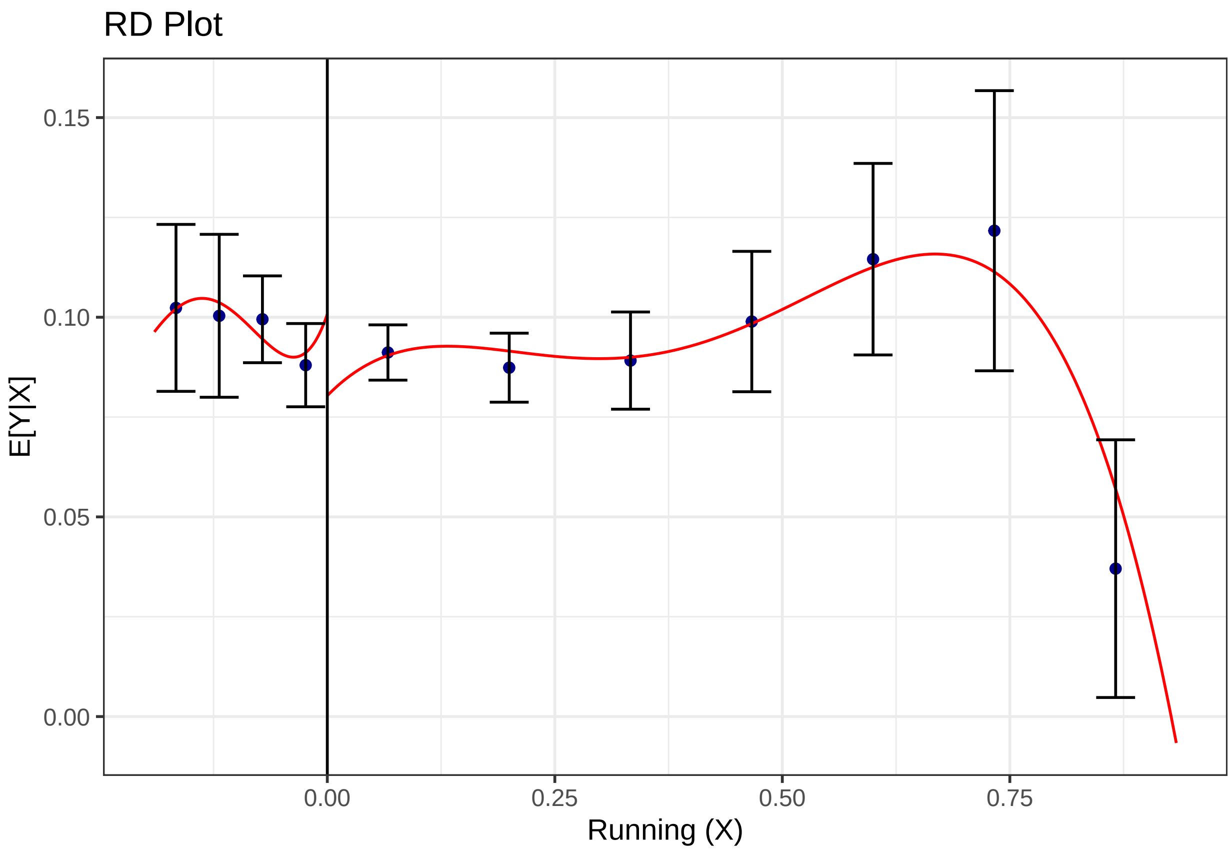

Figure 3: Probability of having a child weighting 2.5Kg or less at birth.

Table 4: RDD estimates. Triangular and Uniform kernels.

Method	Estimate	Std.Error	P-value
Uniform			
Conventional	-0.0453	0.0317	0.1534
Bias-Corrected	-0.0487	0.0317	0.1243
Robust	-0.0487	0.0370	0.1872
Triangular			
Conventional	-0.0387	0.0311	0.2138
Bias-Corrected	-0.0413	0.0311	0.1842
Robust	-0.0413	0.0358	0.2485

Optimal bandwidth was calculated according to the Calonico et al. method as implemented in the R package `rdrobust`. We report bias corrected estimate and variance (Robust), only estimate (Bias-Corrected) and no correction at all (Conventional).

The second stage results show NO EFFECT of Panes program on the probability of giving birth to a low weight child. The result is robust to kernel choice and bias correction.

The first stage relation is the effect of running variable in the probability of getting the treatment, at the cutoff value. This relation is positive and significant according to our results below. This was expected, since Figure 2 showed a relevant positive discontinuity at the cutoff value.

```
## Call: rdrobust
##
## Number of Obs.                22534
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                7404       15130
## Eff. Number of Obs.           2039        1955
## Order est. (p)                   1           1
## Order bias  (q)                  2           2
## BW est. (h)                  0.034       0.034
## BW bias (b)                  0.058       0.058
## rho (h/b)                    0.589       0.589
## Unique Obs.                   6216       11238
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional     0.617     0.027    23.251     0.000     [0.565 , 0.669]
## Bias-Corrected     0.608     0.027    22.931     0.000     [0.556 , 0.660]
##         Robust     0.608     0.031    19.711     0.000     [0.548 , 0.669]
## =============================================================================
```

4. In order to assess the credibility of your empirical strategy, choose a variable which you may argue is predetermined and estimate the effect of program participation at the cutoff as in the previous item. What do you find?

We have chosen the gestational length in weeks, `semgest` as a placebo variable, since it is mainly determined by biological traits of the mother and the PANES program has no influence on these traits. This is a predetermined variable, the cash-transfer program should have no effect on this outcome and this is indeed the case as shown in results below.

```
## Call: rdrobust
##
## Number of Obs.                22534
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 7404       15130
## Eff. Number of Obs.            1768        1697
## Order est. (p)                    1           1
## Order bias  (q)                   2           2
## BW est. (h)                   0.030       0.030
## BW bias (b)                   0.062       0.062
## rho (h/b)                     0.479       0.479
## Unique Obs.                    6216       11238
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional     0.090     0.226     0.397     0.691    [-0.354 , 0.534]
## Bias-Corrected     0.092     0.226     0.405     0.686    [-0.352 , 0.535]
##         Robust     0.092     0.250     0.367     0.714    [-0.398 , 0.581]
## =============================================================================
```

5. Implement a manipulation test for the running variable in your setting. What do you find?

We implemented the test suggested by Cattaneo, Jansson, and Ma (2020) and visually inspecting the plot below we are not able to reject the null hypothesis of no manipulation in the running variable.

A summary of this test is provided below.

```
##
## Manipulation testing using local polynomial density estimation.
##
## Number of obs =       22534
## Model =               unrestricted
## Kernel =              triangular
## BW method =           estimated
## VCE method =          jackknife
##
## c = 0                 Left of c           Right of c
## Number of obs         7404                15130
## Eff. Number of obs    1485                1426
## Order est. (p)        2                   2
## Order bias (q)        3                   3
## BW est. (h)           0.025               0.025
##
## Method                T                   P > |T|
## Robust                -0.0826             0.9341
```

```
## Warning in summary.CJMrddensity(manipulation): There are repeated observations.
## Point estimates and standard errors have been adjusted. Use option
## massPoints=FALSE to suppress this feature.
```

```
##
```

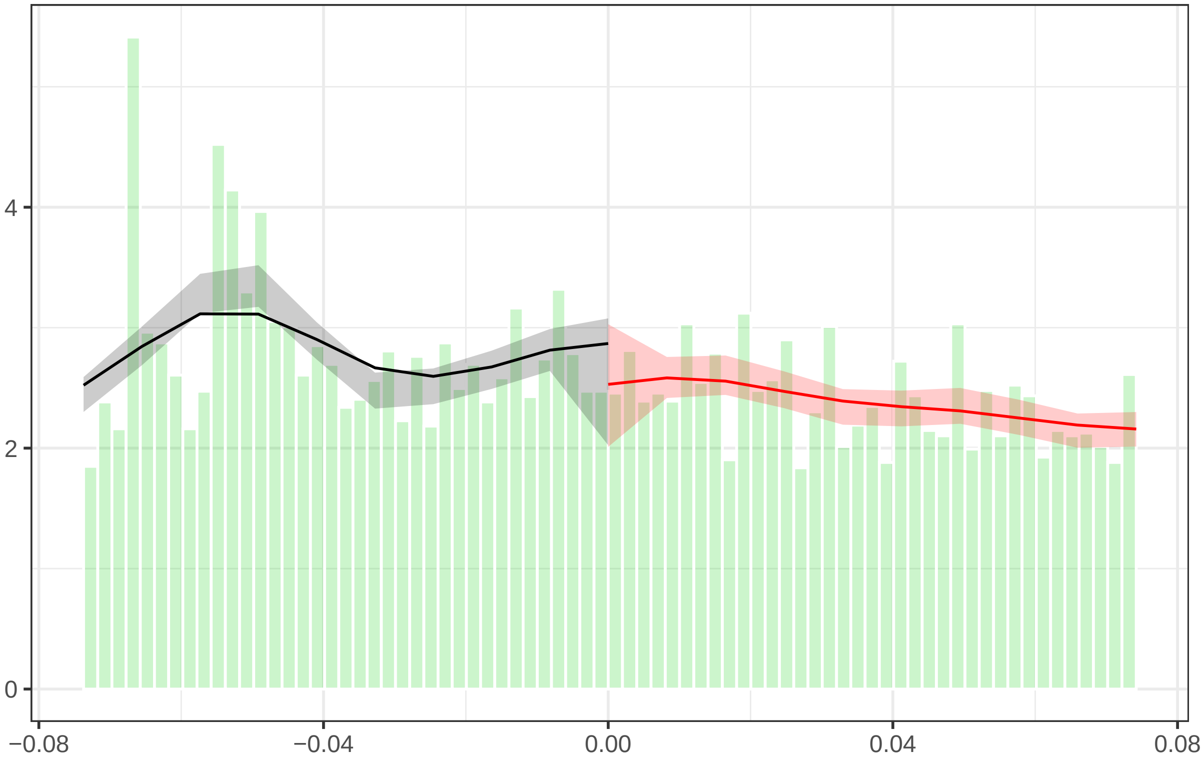

Figure 4: Manipulation test on PANES program

```
## P-values of binomial tests (H0: p=0.5).
##
## Window Length / 2          <c     >=c    P>|T|
## 0.000                       8      12    0.5034
## 0.000                      15      18    0.7283
## 0.001                      21      26    0.5601
## 0.001                      27      31    0.6940
## 0.001                      41      40    1.0000
## 0.001                      52      57    0.7018
## 0.001                      60      70    0.4300
## 0.001                      65      82    0.1868
## 0.002                      74      90    0.2414
## 0.002                      93      99    0.7183
```

6. Do you think there are any potential threats to identification and/or estimation in your context? Can you think of any strategies to circumvent these?

In the fuzzy regression discontinuity design, one of the assumptions for identification is inherited from instrumental variable theory, the exclusion restriction. The PANES program was designed with eligibility (i.e. the running variable) based on a probit model estimated from a baseline survey, and it was target to the poorest 10% of households. The exclusion restriction demands that the running variable does not directly impact the outcome, which in our study is child's low weight at birth. Although, we can easily think of being poor (as the running variable is a proxy for it) is directly linked with the child weight at birth, since it is likely the mother does not eat well and enough, specially for the poorest.

This fact would invalidate our running variable as an instrument, and the way to circumvent this would be to choose another variable from the survey to be the instrument. This new variable should attend both the exclusion restriction and be correlated to the probability of being treated (the relevance condition). The later condition can be met by choosing a variable that was significant in the probit model for eligibility.

Also, for the estimation of a RDD, one must always be very careful about the extrapolation needed when utilizing unities that are not exactly at the cutoff value. Gelman and Imbens (2019) argue that polynomials of order greater than two should not be used in local regressions. High order polynomials are very sensitive to the inclusion of observations and may lead erroneous extrapolation toward the cutoff. We try to avoid this pitfall by using local *linear* regression and choosing a very tight bandwidth. Also

Annex - R Code

```r
#' Homework I - Microeconometrics II
#' Author: Rafael Felipe Bressan
#'
#' Loading libraries
library(tidyverse)
# library(broom)
library(sandwich)
library(lmtest)
library(modelsummary)
library(fixest)
library(AER)
library(dtplyr)
library(rdrobust)
library(rddensity)
library(rdd)
library(foreach)
library(doParallel)


#' Part 1 Instrumental variables
#' Load data
lg_census2010 <- read_csv("II/input/censo_data_family_size.csv")
#' Exploratory analysis
summary(lg_census2010)
#' First child years of education has 16228 NAs. This is our outcome.
lg_census2010 <- lg_census2010 %>%
  filter(!is.na(first_child_years_of_education))

#' Question 1
#' b) Table with average years of schooling by number of children and frequency
#' of number of children. Use sample weight (person_weight)
educ_tab <- lg_census2010 %>%
  group_by(family_number_children) %>%
  summarise(frequency = sum(person_weight),
            avg_school = sum(first_child_years_of_education*person_weight)/frequency) %>%
  select(family_number_children, avg_school, frequency)
educ_tab
#' Total population represented
sum(educ_tab$frequency)
#'
#' c) ATE lower bound
#'
#' ManskiPepper2000 table 1
manski_tbl1 <- lg_census2010 %>%
  select(c(first_child_years_of_education, family_number_children,
           person_weight)) %>%
  group_by(family_number_children) %>%
  summarise(average = weighted.mean(first_child_years_of_education, person_weight),
            prob = sum(person_weight)/sum(lg_census2010$person_weight),
            size = sum(person_weight)) %>%
  arrange(family_number_children)
#' Manski bounds with MTR and MTS
```

```r
ate_ub <- function(data, outcome, treatment, treat_levels, weight) {
  treat_levels <- sort(treat_levels)
  treat_vec <- data %>%
    distinct({{treatment}}) %>%
    arrange({{treatment}}) %>%
    pull()
  if (!all(treat_levels %in% treat_vec))
    stop("All treatment levels must be in the data.")
  # Conditional expectations and probabilities.
  # data must be the manski_tbl1 format
  sum_weights <- data %>% select({{weight}}) %>% pull() %>% sum()
  exp_probs <- data %>%
    select(c({{outcome}}, {{treatment}}, {{weight}})) %>%
    group_by({{treatment}}) %>%
    summarise(average = weighted.mean({{outcome}}, {{weight}}),
              prob = sum({{weight}})/sum_weights,
              size = sum({{weight}})) %>%
    arrange({{treatment}})

  ub <- vector("numeric", length = length(treat_levels))
  n_treats <- nrow(exp_probs)

  for (i in seq_along(treat_levels)) {
    t <- treat_levels[i]
    t_idx <- which(treat_vec == t)
    s <- treat_vec[t_idx - 1]
    sum_t_plus <- exp_probs %>%
      filter({{treatment}} > t) %>%
      summarise(sum(average*prob)) %>%
      pull()
    sum_s_less <- exp_probs %>%
      filter({{treatment}} < s) %>%
      summarise(sum(average*prob)) %>%
      pull()
    exp_y_t <- exp_probs %>%
      filter({{treatment}} == t) %>%
      pull(average)
    exp_y_s <- exp_probs %>%
      filter({{treatment}} == s) %>%
      pull(average)
    prob_t_leq <- exp_probs %>%
      filter({{treatment}} <= t) %>%
      summarise(sum(prob)) %>%
      pull()
    prob_s_geq <- exp_probs %>%
      filter({{treatment}} >= s) %>%
      summarise(sum(prob)) %>%
      pull()
    upper_bound <- sum_t_plus + exp_y_t * prob_t_leq - (sum_s_less + exp_y_s * prob_s_geq)

    ub[i] <- upper_bound
  }
  return(data.frame(treat_levels = treat_levels, upper_bound = ub))
}
#' Estimate of LB
#' Wrong sign. Just checking
ate_lb_est <- ate_ub(lg_census2010, first_child_years_of_education,
                     family_number_children, (2:16), person_weight) %>%
  rename(lb_est = upper_bound)
#' Changing sign They are the same!! Compute upper bound and say it is lower
#' is the same as changing signs, compute UB then change sign back again to
#' lower
#'
#' Bootstrapping for Lower Bound confidence
#' Results in unique households and their weights

lg_household <- lg_census2010 %>%
  select(id_household, household_weight) %>%
  group_by(id_household) %>%
  summarise(household_weight = first(household_weight))

nrep <- 1000
n_sample <- nrow(lg_household)
#' Using foreach
#' Register doParallel backend for parallel computation of bootstrap
cl <- makePSOCKcluster(3)
registerDoParallel(cl)

ate_lb_df <- foreach(i = 1:nrep, .combine = "rbind", .packages = "dplyr") %dopar% {
  boot <- sample(n_sample, replace = TRUE,
                 prob = lg_household$household_weight
  )

  lg_hh_boot <- lg_household[boot, "id_household"] %>%
    left_join(lg_census2010 %>%
                select(id_household, first_child_years_of_education,
                       family_number_children, person_weight),
              by = "id_household")

  lg_hh_boot %>%
    ate_ub(first_child_years_of_education, family_number_children, (2:6),
           person_weight)
}
stopCluster(cl)
gc() # Garbage collector to clean up memory

#' Faceted plot of lower bounds histograms
lb_hist <- ggplot(ate_lb_df, aes(upper_bound, group = treat_levels)) +
  geom_histogram(bins = 20) +
  facet_wrap(~treat_levels, scales = "free") +
  labs(x = "Lower bound") +
  theme_classic()

manski_tbl2 <- ate_lb_df %>%
  group_by(treat_levels) %>%
  summarise(quant025 = quantile(upper_bound, probs = 0.025),
            quant975 = quantile(upper_bound, probs = 0.975)) %>%
  mutate(s_treat = treat_levels - 1) %>%
  left_join(ate_lb_est, by = "treat_levels") %>%
  select(s_treat, treat_levels, lb_est, quant025, quant975)

#’ Question 2
#’ a) Multiple births in second birth (second child are twins)
lg_two_plus_births <- lg_census2010 %>%
  filter(family_number_births >= 2)
n_two_plus_births <- nrow(lg_two_plus_births)
n_second_multiple <- sum(lg_two_plus_births$second_birth_ismultiplebirth,
                         na.rm = TRUE)
#’ b) IV Estimation
iv_est0 <- ivreg(
  first_child_years_of_education~family_number_children
  |second_birth_ismultiplebirth,
  weights = person_weight,
  data = lg_two_plus_births)
cl_vcov0 <- vcovCL(iv_est0, cluster = ~id_household)
iv_est1 <- ivreg(
  first_child_years_of_education~family_number_children+first_child_age+
    first_child_is_girl+first_birth_ismultiplebirth
  |second_birth_ismultiplebirth+first_child_age+first_child_is_girl+
    first_birth_ismultiplebirth,
  weights = person_weight,
  data = lg_two_plus_births)
cl_vcov1 <- vcovCL(iv_est1, cluster = ~id_household)
#’ F-test for instrument relevance
f_test0 <- summary(iv_est0, diagnostics = TRUE,
                   vcov. = vcovCL, cluster = id_household)
f_test0 <- f_test0$diagnostics[’Weak instruments’, ’statistic’]
f_test1 <- summary(iv_est1, diagnostics = TRUE,
                   vcov. = vcovCL, cluster = id_household)
f_test1 <- f_test1$diagnostics[’Weak instruments’, ’statistic’]
#’ c) Weak instrument test
#Code to implement the Olea Pflueger correction:
#FS is
#x= Z\pi +  R \gamma + epsilon

#Arguments are
# edndogenous = character variable with name of dependent variable
# instruments = character vector with name of instruments
# vcov = variance matrix to be used in computing first stage, e.g. vcovCL for cluster robust
# data = data.frame with data
# controls = vector with controls to be included in refression. c() if no controls.
# intercept = shoudl an intercept be included in formulas? Variable named intercept will be included am
#             controls
# weights = vector of weights, if weighted regression is desired
# cluster = if vcov = vcovCL, the name of the cluster variable (defaults to NULL)
# ... = additional arguments, to be passed to vcov function, e.g. degree of freedom correction
olea_pflueger_f <- function(endogenous, instruments, vcov, data, controls = c(),
                            intercept = TRUE, weights = NULL, cluster = NULL, ...)
```

```r
{
  # Early chechk for weights and cluster
  if (!is.null(weights))
    weights <- as.data.frame(data)[, weights]
  if (!is.null(cluster))
    cluster <- as.data.frame(data)[, cluster]

  if (length(controls) > 0)
    data.kept = data[,c(endogenous,instruments, controls)]
  else
    data.kept = data[,c(endogenous,instruments)]

  keep_ind = complete.cases(data.kept)

  data.kept = data.kept[keep_ind,]

  Nobs = nrow(data.kept)

  if (intercept)
  {
    data.kept = cbind(data.kept, "intercept" = 1)
    controls = c(controls, "intercept")
  }

  if (length(controls) > 0)
  {
    # y = as.vector(residuals(lm(as.formula(paste(endogenous, "~ -1 + ", paste(controls,collapse = "+")

    Z = c()

    for (instrument in instruments)
    {
      if(is.null(weights))
        z =  as.vector(residuals(lm(as.formula(paste(instrument, "~ -1 + ", paste(controls,collapse = "
      else
        z =  as.vector(residuals(lm(as.formula(paste(instrument, "~ -1 + ", paste(controls,collapse = "

      Z = cbind(Z, z)
    }

  } else {
    # y = as.vector(data.kept[,endogenous])

    Z = as.matrix(data[,instruments])

  }

  formula.fs = paste(endogenous,"~ -1 +",paste(c(instruments,controls),collapse = " + "))

  if(is.null(weights))
    fs.reg = lm(as.formula(formula.fs), data.kept)
  else
    fs.reg = lm(as.formula(formula.fs), data.kept,
                weights = weights[keep_ind])

  # if(is.null(weights))
  #   fs.reg = lm(y~Z-1) else fs.reg = lm(y~Z-1, weights = weights[keep_ind])

  coefs = fs.reg$coefficients[names(fs.reg$coefficients)%in%instruments]

  if(!is.null(cluster))
    vcov_mat = vcov(fs.reg, cluster = cluster[keep_ind], ...)
  else
    vcov_mat = vcov(fs.reg, ...)

  #Restricting to only instruments
  vcov_mat = vcov_mat[names(fs.reg$coefficients)%in%instruments,names(fs.reg$coefficients)%in%instrument

  if (is.null(weights))
    Q_Z_norm = (t(Z) %*% Z)/Nobs
  else {
    # Q_Z_norm = t(Z)%*%diag(weights[keep_ind])%*%Z/Nobs
    Z_w <- sapply(1:ncol(Z), function(j) weights[keep_ind]*Z[, j])
    Q_Z_norm <- (t(Z_w) %*% Z_w)/Nobs
  }
  F_eff = t(coefs)%*%Q_Z_norm%*%coefs/sum(diag(vcov_mat%*%Q_Z_norm))


  return(list("Nobs" = Nobs, "eff_F" = F_eff))
}


weak_iv_test <- olea_pflueger_f("family_number_children",
                                "second_birth_ismultiplebirth",
                                vcovCL, data = lg_two_plus_births,
                                controls = c("first_child_age",
                                             "first_child_is_girl",
                                             "first_birth_ismultiplebirth"),
                                weights = "person_weight",
                                cluster = "id_household")
#' d) Anderson-Rubin confidence interval
#'
#AR CI

#Arguments are
# outcome = character variable with name of outcome variable of interest
# edndogenous = character variable with name of endogenous variable
# instruments = character vector with name of instruments
# vcov = variance matrix to be used in pooled model.
# data = data.frame with data
# grid_beta = grid over which to perform grid search. Beta is the parameter of
#             interest in the IV regression
# confidence = confidence level for CI (defaults to 0.95)
# controls = vector with controls to be included in refression. c() if no controls.
# intercept = shoudl an intercept be included in formulas? Variable named intercept will be included am
#             controls
# weights = column name in data giving the desired observations weight
# cluster = if vcov = vcovCL, the name of the cluster variable (defaults to NULL)
# ... = additional arguments, to be passed to vcov function, e.g. degree of freedom correction

anderson_rubin_ci <- function(outcome, endogenous, instruments, vcov, data,
                              grid_beta, confidence = 0.95, controls = c(),
                              intercept = TRUE, weights = NULL,
                              cluster = NULL, ...)
{
  # Early chechk for weights and cluster
  if (!is.null(weights))
    weights <- as.data.frame(data)[, weights]
  if (!is.null(cluster))
    cluster <- as.data.frame(data)[, cluster]

  if (length(controls) > 0)
    data.kept = data[,c(outcome, endogenous,instruments, controls)]
  else
    data.kept = data[,c(outcome, endogenous,instruments)]

  keep_ind = complete.cases(data.kept)

  data.kept = data.kept[keep_ind,]

  Nobs = nrow(data.kept)

  if (intercept){
    data.kept = cbind(data.kept, "intercept" = 1)
    controls = c(controls, "intercept")
  }

  #We will pool outcome and endogenous now
  data.pooled = rbind(data.kept, data.kept)

  data.pooled$pool_variable = c(data.kept[,outcome], data.kept[,endogenous])
  data.pooled$variable_indicator = as.factor(c(rep("reduced_form",    Nobs),rep("first_stage",Nobs)))



  #Constructing the formula for regression
  if(length(controls)>0)
    formula = paste("pool_variable ~ -1 + ", paste(paste("variable_indicator", instruments, sep = ":"),c
  else
    formula = paste("pool_variable ~ -1 +", paste(paste("variable_indicator", instruments, sep = ":"),co


  if(is.null(weights))
    pool.model = lm(formula, data.pooled)
  else  {
    pool_weights <- rep(weights[keep_ind], 2)
    pool.model = lm(as.formula(formula), data = data.pooled,
                    weights = pool_weights)
  }
```

```
coefs = pool.model$coefficients

if(!is.null(cluster))
  vcov_model = vcov(pool.model, cluster = rep(cluster[keep_ind],2), ...)
else
  vcov_model = vcov(pool.model, ...)

p1 = grepl(paste("reduced_form", instruments, sep = ":"), names(coefs))
p2 = grepl(paste("first_stage", instruments, sep = ":"), names(coefs))

acc_vec = c()

#Looping over grid
for(beta in grid_beta)
{

  lin_vec = p1 - beta*p2
  #constructing test statistic
  val = (coefs%*%lin_vec)
  vcov_lin = t(lin_vec)%*%vcov_model%*%lin_vec

  ar = val%*%solve(vcov_lin)%*%val

  pvalue = pchisq(ar, 1)

  if(pvalue<= confidence)
    acc_vec = c(acc_vec, T) else acc_vec = c(acc_vec, F)

}

if(sum(acc_vec) == 0)
  return("Confidence set is empty!") else {

    vec_region_start = c()
    vec_region_end = c()
    if(acc_vec[1] == TRUE)
    {
      warning("Lower boundary point was accepted! Perhaps decrease grid lower bound to see what happen
      vec_region_start = grid_beta[1]
    }

    if(acc_vec[length(acc_vec)] == TRUE)
    {
      warning("Upper boundary point was accepted! Perhaps increase grid upper bound to see what happen
      vec_region_end = grid_beta[length(acc_vec)]
    }

    vec_region_start = c(vec_region_start, grid_beta[c(FALSE,diff(acc_vec)==1)]   )
    vec_region_end = c(grid_beta[c(diff(acc_vec) ==-1, FALSE)],vec_region_end)

    # CI.text = paste(paste("[",vec_region_start, ",", vec_region_end, "]"),collapse = " U ")

    return(c(vec_region_start, vec_region_end))
    }
}
#’ beta is estimated to be around -0.4, thus a grid surrounding this value.
grid_beta <- seq(-10, 10, 0.1)
ar_ci <- anderson_rubin_ci("first_child_years_of_education",
                           "family_number_children",
                           "second_birth_ismultiplebirth",
                           vcovCL, lg_two_plus_births, grid_beta,
                           controls = c("first_child_age",
                                        "first_child_is_girl",
                                        "first_birth_ismultiplebirth"),
                           weights = "person_weight",
                           cluster = "id_household"
                           )
ar_ci_txt <- paste0("[", format(ar_ci[1], digits = 4), ", ",
                    format(ar_ci[2], digits = 4), "]")
iv_ci <- coefci(iv_est1)[’family_number_children’,]
iv_ci_txt <- paste0("[", format(iv_ci[1], digits = 4), ", ",
                    format(iv_ci[2], digits = 4), "]")

#’ PART II RDD
#’
#’ Load data and drop unities with NA for treat or running
panes <- data.table::fread("II/input/amarante2016.csv", encoding = "Latin-1",
                           na.strings = "")[!is.na(treat) & !is.na(running)]
# lg_panes <- data.table::fread("II/input/amarante2016.csv", encoding = "Latin-1")

#’ 2) Plots of treat versus running and bajo2500 versus running
rdplot(panes$treat, panes$running, binselect = ’es’, ci = 95)

#’ 3) Local linear regression
#’ Using the rdd package
# rdd_model <- RDestimate(bajo2500~running+treat, data = panes, model = TRUE)
#’ Using rdrobust package
tri_model <- rdrobust(panes$bajo2500, panes$running,
                      fuzzy = panes$treat, kernel = "tri", all = TRUE)
uni_model <- rdrobust(panes$bajo2500, panes$running,
                      fuzzy = panes$treat, kernel = "uni", all = TRUE)
#’ First stage using the bandwidth from fuzzy model
tri_bw <- tri_model$bws["h", ]
tri_first <- rdrobust(panes$treat, panes$running, kernel = "tri", all = TRUE)

# Extract Data Values
table1 <- data.frame(uni_model$coef, uni_model$se, uni_model$pv)
table2 <- data.frame(tri_model$coef, tri_model$se, tri_model$pv)
rdd_met <- rep(rownames(table1), 2)
est_tab <- cbind(rdd_met, rbind(table1, table2))
rownames(est_tab) <- NULL
colnames(est_tab) <- c("Method", "Estimate", "Std.Error", "P-value")

#’ 4) Placebo tests
#’ Gestational length in weeks: semgest
#’ Week first prenatal visit: semprim
```

```r
#' Number of previous pregnancies: numemban
placebo_m1 <- rdrobust(panes$semgest, panes$running,
                           fuzzy = panes$treat, kernel = "tri", all = TRUE)
# placebo_m2 <- rdrobust(panes$semprim, panes$running,
#                            fuzzy = panes$treat, kernel = "tri", all = TRUE)
# placebo_m3 <- rdrobust(panes$numemban, panes$running,
#                            fuzzy = panes$treat, kernel = "tri", all = TRUE)
#' 5) Manipulation test with rddensity
manipulation <- rddensity(panes$running)
manip_plot <- rdplotdensity(manipulation, panes$running)


#' Do not save large dataframes that start with "lg_"
save(list = ls()[!grepl("^lg_.*", ls())], file = "II/input/homework_I.RData")
```

References

Amarante, Verónica, Marco Manacorda, Edward Miguel, and Andrea Vigorito. 2016. "Do Cash Transfers Improve Birth Outcomes? Evidence from Matched Vital Statistics, Program, and Social Security Data." *American Economic Journal: Economic Policy* 8 (2): 1–43.

Andrews, Isaiah, James H Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11: 727–53.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2020. "Simple Local Polynomial Density Estimators." *Journal of the American Statistical Association* 115 (531): 1449–55. https://doi.org/10.1080/01621459.2019.1635480.

Gelman, Andrew, and Guido Imbens. 2019. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics* 37 (3): 447–56.

Manski, Charles F. 2009. *Identification for Prediction and Decision*. Harvard University Press.

Manski, Charles F., and John V. Pepper. 2000. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica* 68 (4): 997–1010. http://www.jstor.org/stable/2999533.

Ponczek, Vladimir, and Andre Portela Souza. 2012. "New Evidence of the Causal Effect of Family Size on Child Quality in a Developing Country." *Journal of Human Resources* 47 (1): 64–106.