SDA 2019 St Gallen: Regression Discontinuity Design and role of government transfers on political support

Filip Mellgren

2019-11-01

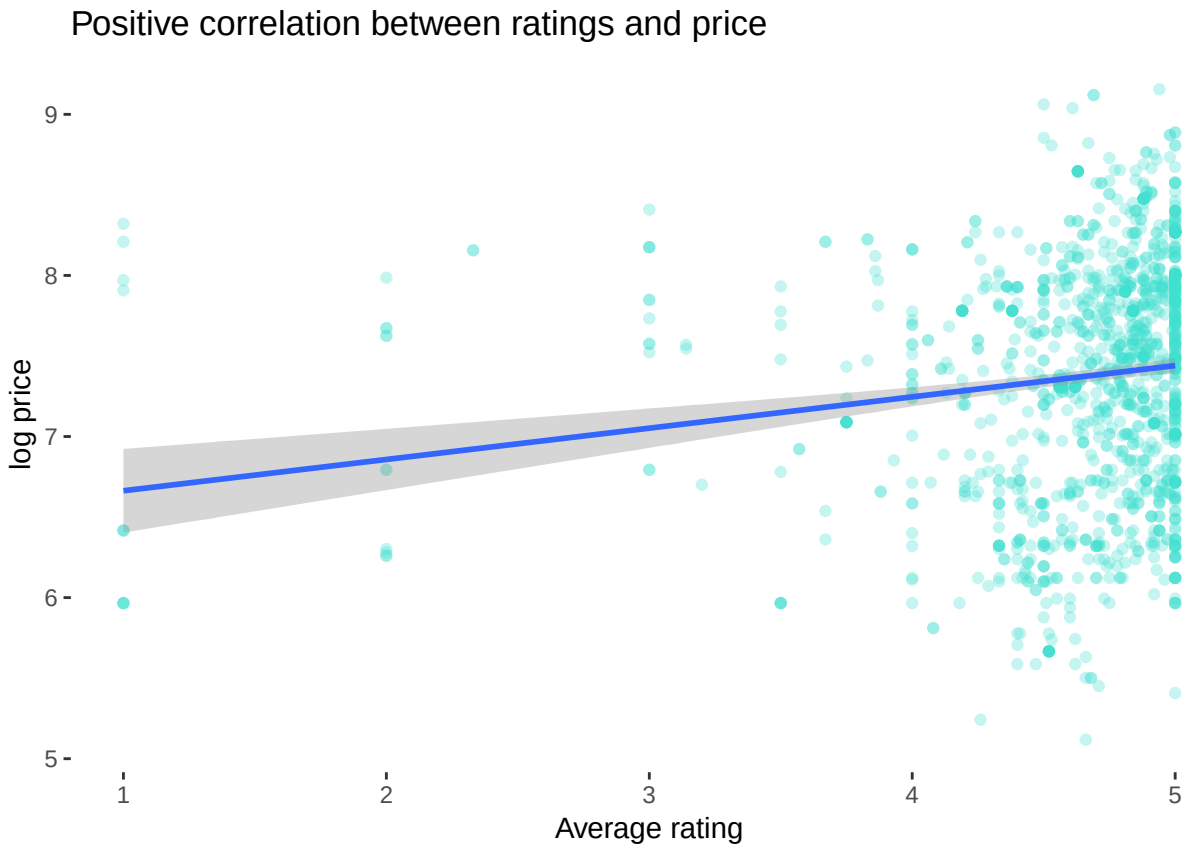
This document presents the topic of regression discontinity design, an econometric tehnnique used to establish causal estimates when the data varies discontinuously around a point due to some institutional feature. The original intention was to do this on Airbnb data collected from the web, however, I proceed with replicating a paper because there were no apparent discontinutiy present in the Airbnb data.

Rating effects on price on Airbnb

To investigate whether average rating has a causal impact on the log price of a listing, it might be possible to compare listings with similar ratings that are displayed differently to users. The idea is that if ratings are rounded (which they turned out not to be presently, but may have been so in the past when the data was taken), then an institutional feature creates variation that is as good as exogenous since ratings on either side of the rounding cutoff will be similarly rated and not expected to have any other systematic differences.

```
df <- import("chic_aug_07.csv")
df %>% mutate(PriceX = gsub("[^0-9.]", "", df$price_string),
              PriceX = as.numeric(PriceX),
              dist_t = avg_rating - round(avg_rating/0.5)*0.5,
              treat = if_else(dist_t == 0, 1, 0)) -> df

df %>% filter(avg_rating > 0) %>% ggplot(aes(x = avg_rating, y = log(PriceX))) +
  geom_point(alpha = 0.3, color = "turquoise") + geom_smooth(method = "lm", formula = y~x) +
  labs(x = "Average rating", y = "log price", title = "Positive correlation between ratings and price") +
  theme(
        panel.background = element_rect(fill = "transparent",colour = NA),
        plot.background = element_rect(fill = "transparent",colour = NA)
        )
```

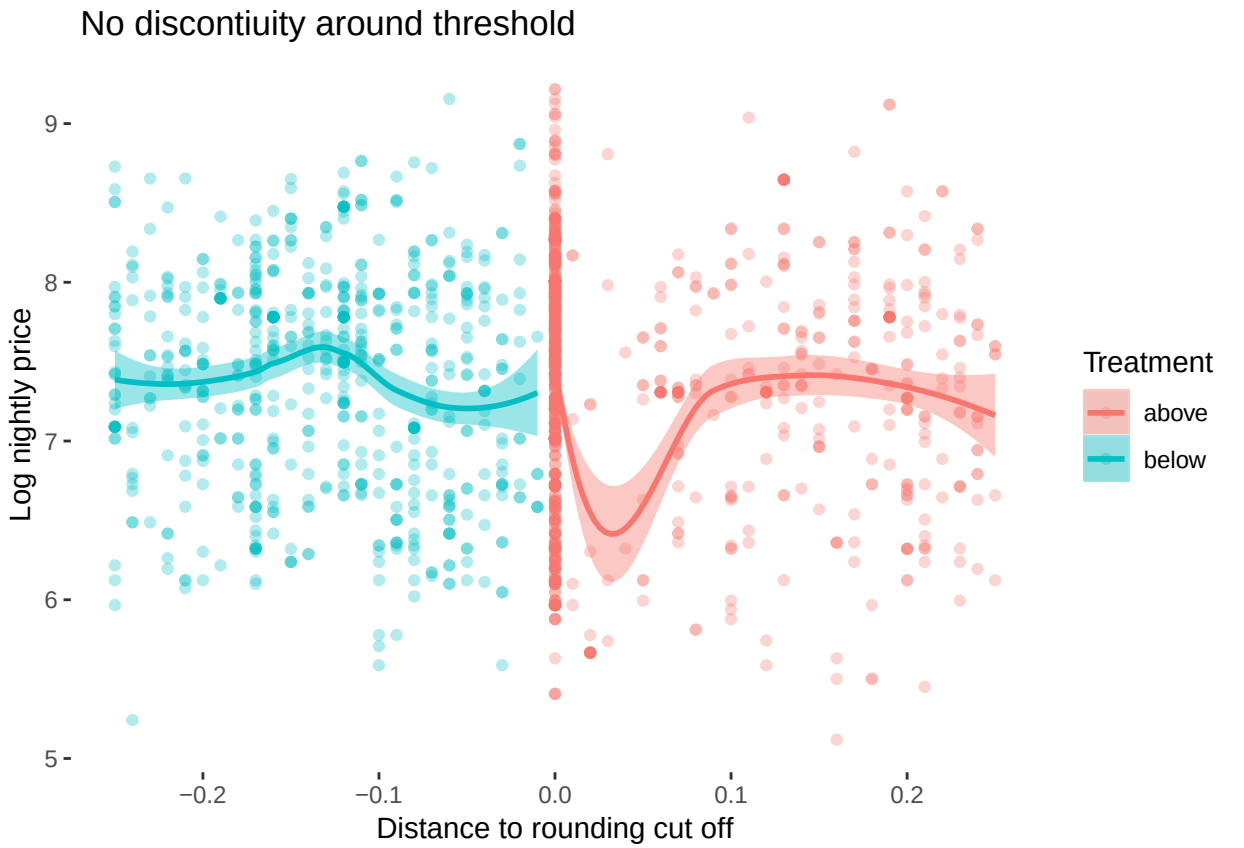



```
ggsave("images/correlation.png", bg = "transparent")
```

There is a correlation present in the data. Let's investigate whether we can use the techniques intended for this analysis to uncover any causal effect.

```
# "Discontinuity plot"

# Rating, log price relationship
df %>% mutate(above = if_else(dist_t >= 0, dist_t, NaN),
                      below = if_else(dist_t < 0, dist_t, NaN)) %>%
  gather(above, below, key = "Treatment", value = "dist_to_cutoff") %>%
  ggplot(aes(x = dist_to_cutoff, y = log(PriceX),
             color = Treatment, fill = Treatment)) +
  geom_smooth(method = loess, span = 0.75) +
  geom_point(alpha = 0.3) +
  labs(title = " No discontiuity around threshold",
       x = "Distance to rounding cut off", y = "Log nightly price") +
  theme(
    panel.background = element_rect(fill = "transparent",colour = NA),
    plot.background = element_rect(fill = "transparent",colour = NA))
```




```
ggsave("images/airbnb_rd.png", bg = "transparent")
```

The plot above strongly suggests there is no such relatioship. I therefore choose to proceed with an already published paper that I know is a meaningful case in point to show what RDD is about.

New analysis, replicating Manacorda, Miguel, Vigorito (2009)

The paper by Manacorda, Miguel, and Vigorito (2011) can be accessed from here: http://emiguel.econ.berkeley.edu/assets/miguel_research/17/_WorkingPaper___Government___Transfers___and___Political_Support.pdf

Abstract of the paper:

> "We estimate the impact of a large anti-poverty cash transfer program, the Uruguayan PANES, on political support for the government that implemented it. Using the discontinuity in program assignment based on a pre-treatment eligibility score, we find that beneficiary households are 11 to 14 percentage points more likely to favor the current government relative to the previous government. Political support effects persist after the program ends. A calibration exercise indicates that these persistent impacts are consistent with a model of rational but poorly informed voters learning about politicians' redistributive preferences."

It strives to establish causal effects of government benefits on government support. The following is noted regarding the current academic debate about the issue:

- There is an empirically established relationship that voters respond to macro economic trends.
- At the same time, there are econometric concerns about endogeneity and small, aggregated samples.
- The true, causal, magnitude is unknown.

In their paper, the authors wishes to establish a causal link between political support of the incumbent party and government transfers. The setting is such that families receive a government program called "PANES" if their predicted income is above a certain threshold. The government bases its support on a prediction because reported income may suffer from being unstable following a major crisis in the country in the foremath of the event, hard to verify claims as the target often worked informally, and multiple variables help mitigate strategic misreporting.

In summary the Panes program looked had the following defining features:

- It was a temporary cash transfer and food voucher following an economic downturn in the year 2002
- Recipients were households with low predicted income "based on a large number of pre-treatment covariates".
  - following crisis, many had an unstable income, so present was bad precitor of permanent income
  - hard to verify income claims as the target populaiton of worked informaly
  - multiple variables helps against strategic misreporting
- Households around the predicted income threshold for receiving Panes were "surveyed and asked a series of questions including their support for the current government, and a second similar follow-up survey took place the following year".

```
df_mccrary <- import("Government Transfers_Replication/mccrary.dta")
df_reg_panes <- import("Government Transfers_Replication/reg_panes.dta")
```

```
# Create variables indicating whether above or below threshold
# Later cleans the h_89 variable
df_reg_panes %>% mutate(above = if_else(ind_reest > 0, ind_reest, NaN),
                        below = if_else(ind_reest < 0, ind_reest, NaN),
                        gov2007 = case_when(
                          h_89 == 9 ~NaN,
                          h_89 == NA ~NaN,
                          TRUE ~ (h_89-1)/2)
                        ) -> df
```

Descriptive Statistics

To begin the analysis, I establish some observational facts by estimating a regression using OLS. Of course, these results will contain endogeneity bias but will nonetheless be useful as a benchmark. Potential reasons for endogeneity might include:

- Recipients already leaned towards one political side.
- Recipients respond they approve of the president in order to not loose benefits
  - However, Uruguay isn't a corrput country, hence unlikely.

```
# Simple regression, approval status on approved for PANES
ols_fit <- lm(gov2007 ~ aprobado + ind_reest, data = df)
stargazer(ols_fit, type = "html")
```

Dependent variable:

gov2007

aprobado

0.166***

(0.035)

ind_reest

0.386

(1.466)

Constant

0.686***

(0.020)

Observations

2,089

R2

0.039

Adjusted R2

0.038

Residual Std. Error

0.388 (df = 2086)

F Statistic

42.486*** (df = 2; 2086)

Note:

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

Overall, there is a negative relationship between predicted income and government support. Likely reflecting party difference in support among various income groups.

Exogenous variation

To escape from the endogeneity present in the previous analysis, we can think of predicted income as a variable with some exogenous variation around the threshold of being assigned the government support. The basic idea is that households on either side of the threshold will be similar in all characteristics but their treatment status.

Falling slightly to the right means some indicators of income predicts the household is above the threshold whereas falling on the left side of the threshold means the predcitors state that the household is poor enough to receive the benefits. However, except for extremely modest variation in the income indicators, predicted income may be deemed as something varying almost entirely randomly.

If there happens to be a large discontinuity in government support around the specified threshold, the effect can plausibly be attributed to treatment status as nothing else may be expected to vary around the somewhat arbtirary threshold of predicted income.

There are, however, a few issues that must be dealt with. First, if certain households (or government officials) anticipate that they may be close to the threshold, they might try to affect the measure predicted income so as to move themselves to the left and thus increase their chance of receiving the government benefit resulting in a potnetially biased sample where the treated population is skewed towards households capable of affecting the outcomes of the government transfer, A second issue relate to potential confounding variables that also vary around the threshold. For example, if it turns out that predicted income was also used for some other government program, then it'd be unclear from what program potential effects adhere. Another issue might be if the metric varies around the threshold based on some observable characteristics, then effects could depend on this characteristic rather than treatment status.

These issues, however, can be checked. The first issue implies a skewed density distirbution around the threshold whereas the second needs to make the case that this variable exists and then show a discontinuity

precisly around the threshold. In practice, it is hard to come up with such variables and especially in this case because predicted income is a construct of several variables and unlikely to be used elsewhere.

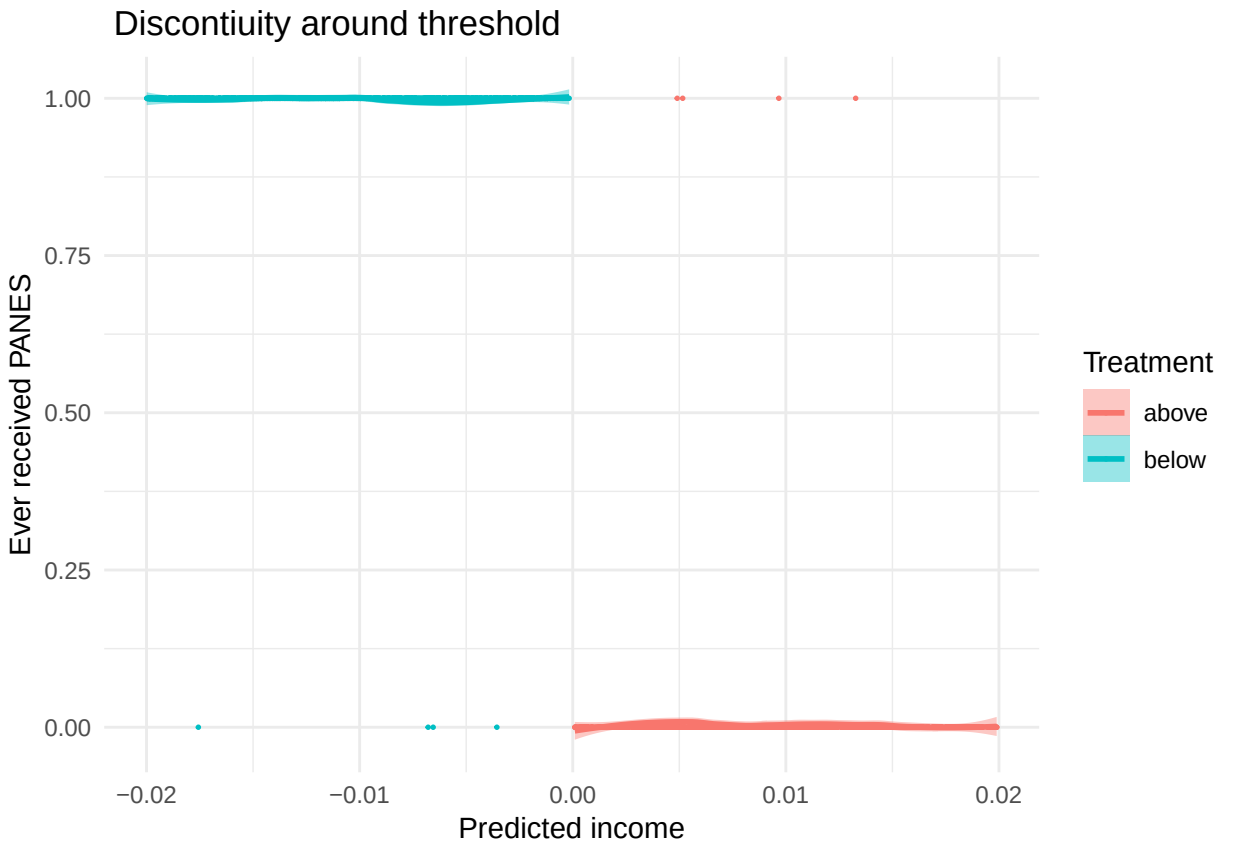
A final concern partains to whether there is a discontinuity or just a sharp (continuous) decline. This will be an issue if the so called "bandwidth" is wide (see the three images below) and influenced by observations far from the cutoff that makes a continuous function appear discontinuous.

There is a connection between a regression discontinuity estimates and instrumental variables (in fact, a fuzzy RDD is an IV estimation, Hahn, Todd, and Van der Klaauw (2001) was first with this interpretation). I therefore proceed by presenting the analysis as if it was an IV together with the rich set of intuitive graphs present in an RDD estimation.

First stage

The first stage in an IV study considers the strength of the exogenous instrument z (referred to as the "instrument"") in predicting the endogenous variable of interest x. Ideally, the instrument is a strong predictor of the endogenous variable to avoid issues related to weak variable bias. See Bound, Jaeger, and Baker (1995) for an explanation of the problem with weak instruments. Other than establishing a tight link between the two variables, the first stage is simply a mechanic check that needs to be done. It's importance is measured by an F statistic which is ideally larger than 10.

```
# First stage
df %>% gather(above, below, key = "Treatment", value = "Predicted_income") %>%
  ggplot(aes(x = Predicted_income, y = aprobado,
             color = Treatment, fill = Treatment)) +
  geom_smooth(method = "loess", span = 0.5) +
  geom_point(size = 0.3) +
  theme_minimal() +
  labs(title = " Discontiuity around threshold",
       x = "Predicted income", y = "Ever received PANES") +
    theme(
        panel.background = element_rect(fill = "transparent",colour = NA),
        plot.background = element_rect(fill = "transparent",colour = NA)
        )
```




```
ggsave("images/first_stage.png", bg = "transparent")
```

As can be seen in the figure above, there is a clear relationship between the two. Namely that when predicted income changes status, so does PANES status. This is intuitively clear because the government is only handing out the PANES to eligible households (those poor according to predicted income). There are a few exceptions but overall, the relationship is very strong which in the words of RDD can be referred to as a "Sharp RDD".

Reduced form, support for government

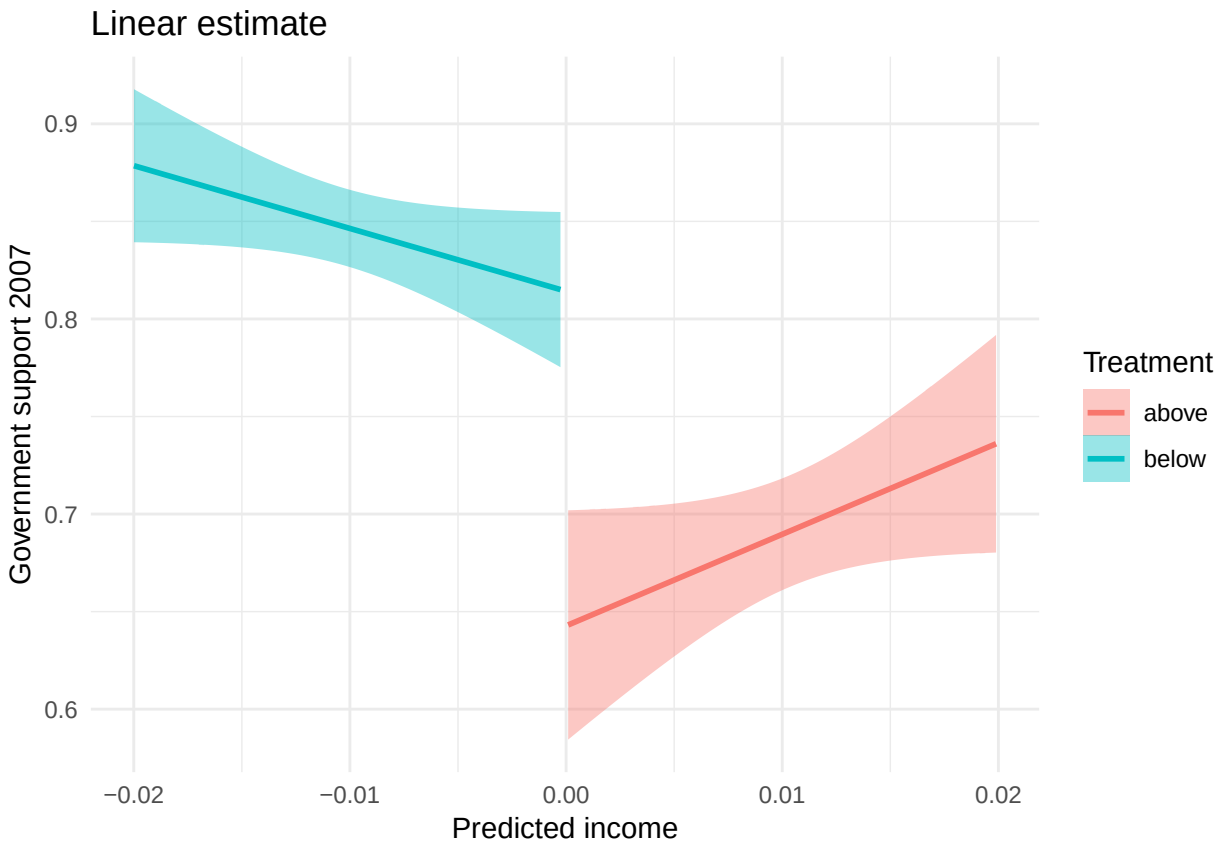
The reduced form can also be interpreted as the intention to treat effect, people on one side of the cutoff were intended to receive PANES and the following graphs show how people intended to receive panes vary their support for the government. The ITT effect is the difference between the graphs around the cutoff.

An important hyperparameter to consider is what bandwidth is selected for the analysis. As can be seen, the local linear regression estimates (the curvy graphs) produce noisier estimates than does the linear relationship. The reason is that they are locally estimated using less data. In principle it is good to establish the causal effects with an as small interval as possible as this leads to less bias. Unfortunately, it also increases variance which is why there has to be a trade off between the two. I show three variations to give a feel for the sensitivity.

```
# Reduced form: support for govt - first wave
# Lower span is less biased, but a noisier estimate.

df %>%
  gather(above, below, key = "Treatment", value = "Predicted_income") %>%
  ggplot(aes(x = Predicted_income, y = gov2007,
```

```
            color = Treatment, fill = Treatment)) +
  geom_smooth(method = lm) +
  theme_minimal() +
  labs(title = "Linear estimate",
       x = "Predicted income", y = "Government support 2007") +
    theme(
        panel.background = element_rect(fill = "transparent",colour = NA),
        plot.background = element_rect(fill = "transparent",colour = NA)
        )
```

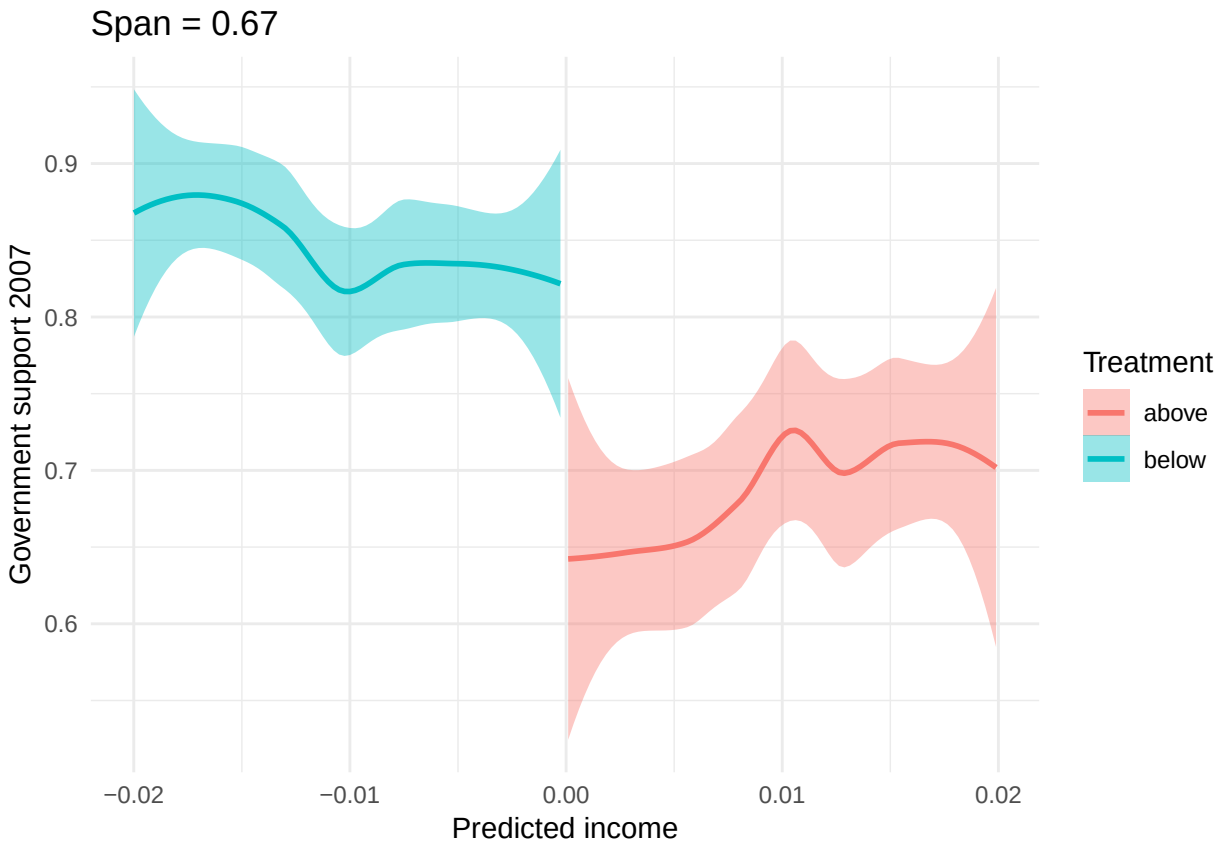



```
ggsave("images/reduced_line.png", bg = "transparent")


df %>%
  gather(above, below, key = "Treatment", value = "Predicted_income") %>%
  ggplot(aes(x = Predicted_income, y = gov2007,
             color = Treatment, fill = Treatment)) +
  geom_smooth(method = "loess", span = 0.33) +
  theme_minimal() +
  labs(title = "Span = 0.33",
       x = "Predicted income", y = "Government support 2007") +
    theme(
        panel.background = element_rect(fill = "transparent",colour = NA),
        plot.background = element_rect(fill = "transparent",colour = NA)
        )
```


```
ggsave("images/reduced_0.33.png", bg = "transparent")

df %>%
  gather(above, below, key = "Treatment", value = "Predicted_income") %>%
  ggplot(aes(x = Predicted_income, y = gov2007,
             color = Treatment, fill = Treatment)) +
  geom_smooth(method = "loess", span = 0.67) +
  theme_minimal() +
  labs(title = "Span = 0.67",
       x = "Predicted income", y = "Government support 2007") +
    theme(
        panel.background = element_rect(fill = "transparent",colour = NA),
        plot.background = element_rect(fill = "transparent",colour = NA)
        )
```




```
ggsave("images/reduced_0.67.png", bg = "transparent")
```

Iv regression

In this part, I formally estimate a model already shown in the graphs based on chapter 6 in Angrist and Pischke (2008).

The model is as follows:

- First stage: $D_i = \gamma_0 + \gamma_1 x_i + \pi T_i + \gamma_2 x_i T_i + \xi_{1i}$
- Reduced form, ITT: $Y_i = \mu + \kappa_1 x_i + \rho\pi T_i + \kappa_2 x_i T_1 + \xi_{2i}$
- Second stage: $Y_i = \alpha + \beta_1 x_1 + \rho D_i + \beta_2 x_i D_i + \eta_i$

Where:

- $T_i = 1(x_i \geq x_0)$ is an indication of the point of discontinuity in predicted income.
- D_i is treatment status that we predict in the first stage.
- Y_i is government approval
- x_i is predicted income, the running variable.

```
df <- df %>% mutate(T_i = if_else(ind_reest >= 0, 1, 0))

iv1 = ivreg(gov2007 ~ ind_reest*aprobado | ind_reest*T_i , data = df)

iv_robust <- robust.se(iv1) # robust s.e.
```

[1] "Robust Standard Errors"

```
stargazer(iv_robust, ols_fit, type = "html")
```

Dependent variable:	
gov2007	
coefficient	
OLS	
test	
(1)	(2)
ind_reest	
4.735*	0.386
(2.586)	(1.466)
aprobado	
0.174***	0.166***
(0.038)	(0.035)
ind_reest:aprobado	
-7.878**	
(3.106)	
Constant	
0.642***	0.686***
(0.031)	(0.020)
Observations	
2,089	
R2	
0.039	
Adjusted R2	
0.038	
Residual Std. Error	
0.388 (df = 2086)	
F Statistic	
42.486*** (df = 2; 2086)	

Note:

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

It is noted that the estimated treatment effect is of about the size indicated by the figures (0.17). This is larger than what the authors report, most likely due to the authors having done some additional data cleaning I've omitted here for brevity and focus on the RDD part. The intercept is larger than 0.5, meaning that more people indicated that they approve of the current president than did not. The interaction term is negative, capturing the negative slope observed to the left in the figures.

```
iv_summary <- summary(iv1, diagnostics = TRUE, vcov = sandwich) # robust s.e. using sandwich
```

In the paper, the authors additionally allow for varying polynomials of order 0, 1, and 2 (columns 1, 2, and 3) on both sides of the threshold (I only report of order 1 here) and also include varying socioeconomic covariates (columns 4, 5, 6) with which the results remain robust. This is precisely what we would expect if the covariates don't also have a discontiuity around the cut off.

Note a few nice properties of the diagnostics, the Wu Hausman statistic is insignificant meaning that it cannot be rejected that OLS and IV estimates are equally consistent. Hence, OLS estimates might have been preferable since they are more efficient. This is also confirmed by the similarity of the point estimates in the table above. Additionally, we have that the first stage is fantastic with a p value on the order of less than computer precision .

```
iv_summary
```

Call: ivreg(formula = gov2007 ~ ind_reest * aprobado | ind_reest * T_i, data = df)

Residuals: Min 1Q Median 3Q Max -0.8779 0.1219 0.1529 0.2713 0.3892

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.64177 0.03145 20.403 < 2e-16 ***ind_reest 4.73521 2.58634 1.831 0.0673 .***
aprobado 0.17375 0.03842 4.523 6.45e-06 ind_reest:aprobado -7.87772 3.10608 -2.536 0.0113 *

Diagnostic tests: df1 df2 statistic p-value
Weak instruments (aprobado) 2 2085 1.534e+04 < 2e-16 ***Weak instruments (ind_reest:aprobado) 2 2085 1.099e+05 < 2e-16*** Wu-Hausman 2 2083 8.809e+00 0.000155 *** Sargan 0 NA NA NA
— Signif. codes: 0 '*** 0.001 **' *0.01* ** 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3876 on 2085 degrees of freedom Multiple R-Squared: 0.0421, Adjusted R-squared: 0.04072 Wald test: 28.27 on 3 and 2085 DF, p-value: < 2.2e-16

Other variables varying around the threshold

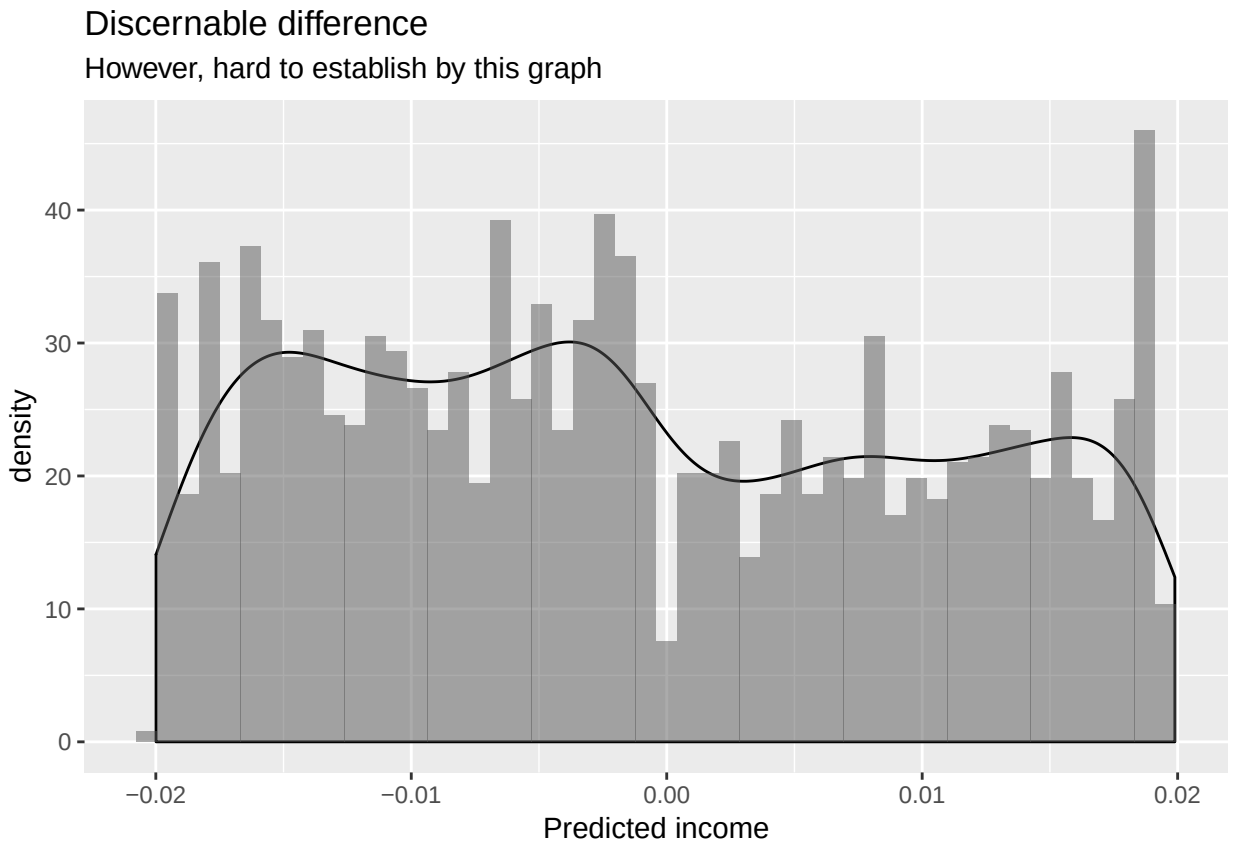
Density around threshold

I begin by showing the density of the running variable around the cutoff to discover potential gaming or self selection around the threshold. If there is gaming, the bar just to the right of the threshold will be "unnaturally" low and the bar to the left will be unaturally high, reflecting that people were bumped to one side because of incentives to do so. If this is the case, the sample will not have been random as some part of the population will be able to self select into treatment. Judging form the graph, it is possible that there is gaming towards the left, where the recipients would just barely qualify for PANES. However, it is not clear cut that this is the case. Fortunately, there are newer methods established to help identify whether there is self selection around the threshold.

Note also that there are few around zero, this likely has to do with the extra survey that ocurred for people close to the threshold.

```
df %>%
  ggplot(aes(x = ind_reest)) +
```

```
  geom_density(trim = TRUE, alpha = 0.4) +
  geom_histogram(aes(y = ..density..), bins = 50, alpha = 0.5) +
  labs(x = "Predicted income",
       title = "Discernable difference", subtitle = "However, hard to establish by this graph")
```



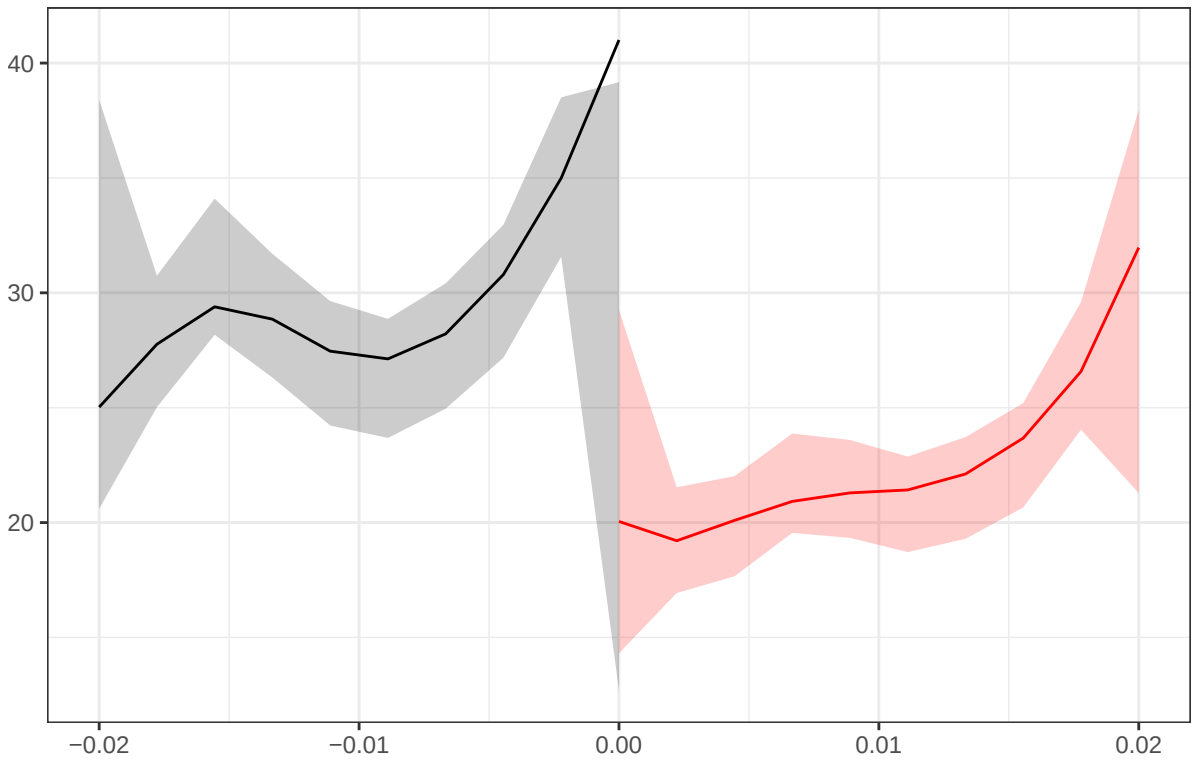

```
ggsave("images/density.png", bg = "transparent")
```

```
## Saving 6.5 x 4.5 in image
```

I proceed by plotting a local polynomial fit of the density. This plot estimates a local polynomial of the density and is based on the paper by Cattaneo, Jansson, and Ma (2019). The idea is that the polynomial fit doesn't suffer from the same boundary bias as do the kernel estimate above. I used the deafult options for bandwidth as calculated by the package for simplicity.

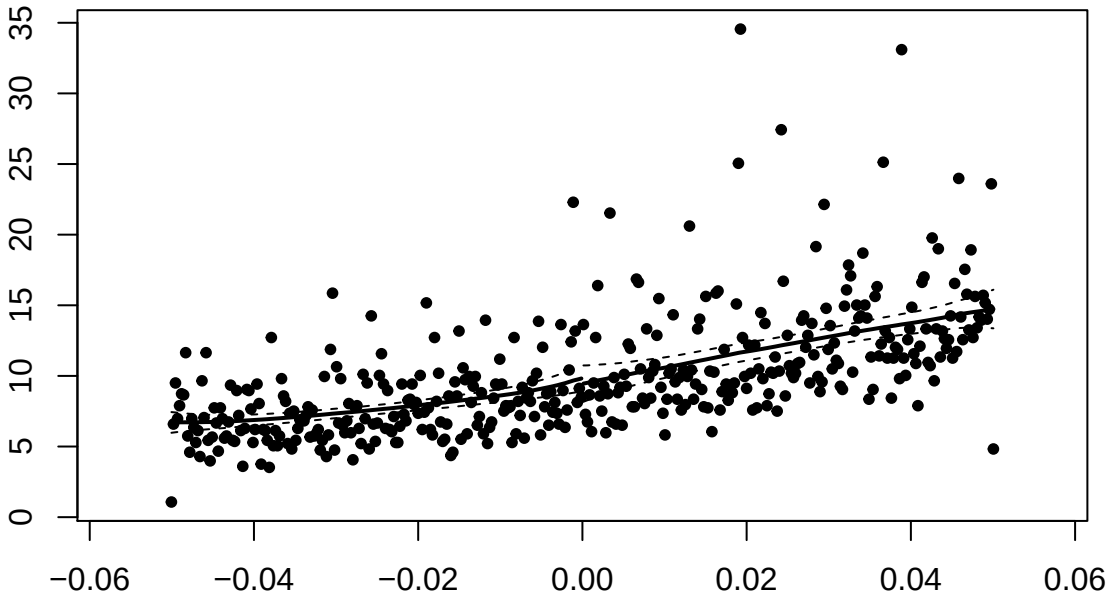
```
rdd <- rddensity(X = df$ind_reest)

# Plot
local_poly_fit <- rdplotdensity(rdd, df$ind_reest, plotRange = c(-0.02, 0.02))
```



The graph above puts doubt on whether there was self selection going on as indicated before; the point estimate of the non recipient population is within the confidence interval and there is significant overlap. To statistically test whether there is a problem, I proceed with the McCrary test.

```
DCdensity(df_mccrary$ind_reest, 0, bin = NULL, bw = NULL, verbose = FALSE, plot = TRUE, ext.out = FALSE
```




```
## [1] 0.1295961
```

The McCrary test ends up being insignificant at $p = 0.13$. This means we cannot reject the null hypothesis that there is no discontinuity around the density of the running variable, which is good news for the analysis as this indicates there were no gaming going on, pushing people to either side of the threshold, so people on both sides can be said to be relatively similar. This, along with the effect sizes estiamted earlier, leads to a

conclusion that the program likely helped the incumbent party.

Conclusion

It appears as if there is a large causal effect of the government program on government support. This is confirmed by the IV estimates of the RDD and robustness checked by running gaming checks. It could still be the case that a covariate varies around the same threshold and is the driver of the effect. However, this seems unlikely as the running variable is an econometric prediction of reported income that was made up of several variables. That one would be discontinuous around the same value as predicted income needs a pervasive case.

The main remaining concern is that reported approval of the president is not a good measure of voting intentions. As was noted earlier, Uruguay is not a corrupt country so respondents had no incentive to report support when they don't support the president. Additionally, the parliamentarianism of Uruguay makes it likely that approval of the president is strongly related to voting intentions of the president's party.

A final note of caution is to interpret the point estimate in this report carefully as the point estimates don't correspond to those of the authors, despite using the same data set. As was noted, this probably boils down to me omitting some data cleaning that the authors may have done to improve the accuracy of reported estimates.

References

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

Bound, John, David A Jaeger, and Regina M Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90 (430): 443–50.

Cattaneo, Matias D, Michael Jansson, and Xinwei Ma. 2019. "Simple Local Polynomial Density Estimators." *Journal of the American Statistical Association*, no. just-accepted: 1–11.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69 (1): 201–9.

Manacorda, Marco, Edward Miguel, and Andrea Vigorito. 2011. "Government Transfers and Political Support." *American Economic Journal: Applied Economics* 3 (3): 1–28.