



PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE ECONOMÍA & ADMINISTRACIÓN
INSTITUTO DE ECONOMÍA

PROBLEM SET # 4

Jose Carlo BERMÚDEZ

jcbermudez@uc.cl

ECONOMETRIC THEORY II

PROFESSOR: TOMÁS RAU

JULY 1ST, 2024

Summary

This report includes my solution to the problem set # 4 for the Econometric Theory II graduate course. The first section of the report exercises randomization. The second part of the document implements causal analysis using Instrumental Variables (IV). The third section addresses Regression Discontinuity (RD). The replication package is available at my [GitHub account](#).

Contents

1	RANDOMIZATION	2
1.1	OLS	2
1.2	MATCHING	3
1.3	ALTERNATIVE EMPIRICAL STRATEGY	6
1.4	ATE ESTIMATION	9
1.5	COMPARISSON BETWEEN ESTIMATORS	11
2	INSTRUMENTAL VARIABLES	12
2.1	OLS REGRESSIONS	12
2.2	DISCUSSING THE INSTRUMENTS	13
2.3	2SLS vs OLS	14
2.4	WEAK INSTRUMENTS TEST	15
2.5	OVERIDENTIFICATION TEST	16
2.6	HAUSMAN TEST	18
2.7	COMMENTS FROM ALBOUY	18
2.8	RESULTS FROM ALBOUY	19
3	REGRESSION DISCONTINUITY	20
3.1	EMPIRICAL STRATEGY	20
3.2	IDENTIFICATION ASSUMPTIONS	20
3.3	LOCAL LINEAR REGRESSION	22
3.4	REPLICATING FIGURES	23
3.5	REPLICATING RD RESULTS	25
3.6	CHANGING THE BANDWIDTH ESTIMATION	25
3.7	USING IK METHOD	26
A	Appendix	29
A.1	ADDITIONAL FIGURES	29

1 RANDOMIZATION

Para mejorar el aprendizaje y desempeño de los estudiantes, los gobiernos de todo el mundo financian una gran variedad de insumos del proceso educativo: textos, libros, material didáctico, etc. El objetivo de este ejercicio es evaluar el efecto de la entrega de uno de estos insumos en el aprendizaje de los estudiantes, medido mediante pruebas estandarizadas (tipo SIMCE). El programa a evaluar corresponde a la entrega de “flip charts” en Kenya. Los flip charts son carteles grandes con elementos didácticos que el profesor puede ir rotando para usar de apoyo en su clase. Usted cuenta con la base de datos `retro-data-kremer.dta`, base con datos de una Organización No Gubernamental (ONG) que distribuye los flip charts en Kenya. La ONG le cuenta que los flip charts se asignan según una regla que incluye los siguientes elementos: (1) capacitación a los profesores, (2) si la sala de clases es cubierta, (3) si el techo tiene goteras, (4) el número de escritorios por alumno, (5) el número de libros por niño y (6) el tamaño de la sala de clases.

1.1 OLS

Estime un modelo mediante MCO para calcular la correlación entre tener algún flip chart (any-flip) y el puntaje en pruebas estandarizadas (nmse). ¿Por qué podría esta correlación no representar el impacto de recibir un flip chart? Mencione al menos 2 ejemplos concretos.

Results are displayed in [Table 1](#). We get that students from schools receiving any flip chart face higher grades by 0.13 standard deviations (since the outcome variable is scaled) compared to those without any flip chart. This coefficient turns out to be statistically significant at any level when we do not include clustered standard errors at the school level –column (1)–, but the significance disappears when clustering is included. I considered both specifications because the data shows overlapping in the sense that all classrooms in the same school are completely treated or not.

Table 1: OLS ESTIMATES

	(1)	(2)
$\mathbb{1}(\text{Flip chart} = 1)$	0.127*** (0.009)	0.127 (0.126)
Constant	-0.035*** (0.005)	-0.035 (0.042)
N	54,782	54,782

Note: This table reports OLS estimates. Column (1) includes robust standard errors, while column (2) includes clustered standard errors at the school level. No controls are included.

Let us consider the case when standard errors are clustered at the school level. We could assume that this implies that flip charts have a null impact on students’ scores, but these are misleading conclusions since we can not claim the coefficient is an unbiased estimate of the causal effect of the program. Some reasons for that are enlisted next:

- **Selection bias.** The Non-Profit Organization (ONG) assigns the flip charts according to several observable traits (professor training, classroom characteristics, etc) that might be correlated to other covariates that also affect student performance like school infrastructure or teachers motivation. We do not have any evidence suggesting that this example case constitutes formal randomization, which would imply that all schools face the same probability of receiving the flip charts. Randomization is also important because all treated and non-treated units would be equal in characteristics at baseline except for the use of flip charts per se. In the absence of randomization is hard to claim that an OLS model for the difference in means captures any causal effect. Thus, the estimates above are biased.

- **Omitted variables/confounders.** There might exist other factors affecting both the probability of receiving a flip chart and the student's performance like the managerial capabilities of school deans or any other source of additional funding from the ONG, like any mentorship program for students with special learning capabilities. Suppose these confounders have a statistically significant effect on the normalized score and we do not control for them. In that case, the correlation between flip chart assignments and standardized tests might reflect the impact of managerial skills or additional funding and not necessarily the effect of using flip charts.
- **Spillover effects.** Let us imagine that the assignation of the flip charts is not made at the school level, which would become a plausible threat to the Stable Unit Treatment Value (SUTVA) assumption as within the school some classrooms would receive the flip charts while others don't. Eventually, teachers from classrooms that didn't receive any flip charts in the same school would adopt this material as a teaching method, thus, making it difficult to disentangle between the treatment effects of using flip charts and non-intended effects.

1.2 MATCHING

Estime el impacto de recibir flip charts mediante un modelo de matching usando las 6 variables mencionadas más arriba. Escoja el método de matching que según usted mejor se ajusta a este escenario. Muestre el balance post matching para agregar robustez a su estimación.

I will estimate treatment effects by implementing nearest-neighbor (NN) matching ([Abadie & Imbens, 2006](#)). This method matches treated units with untreated peers according to a set of observable covariates, this is why is also known as “matching in characteristics”. Every treated unit will have a comparable clone that mirrors its characteristics in a way such that perfect balance is guaranteed, thus, removing any selection bias. However, implementing nearest-neighbor matching requires to selection of the number of neighbors. The best number is often chosen between 3 and 16, but there is no robust theoretical basis for this so we rely on a rule of thumb instead. I estimate Average Treatment Effects (ATE) and Average Treatment on the Treated (ATT) using 1, 3, 6, and 10 NN with the inverse diagonal sample covariate covariance as the distance metric.

Results are reported in [Table 2](#). Two takeaways are drawn from these results. First, matching estimates suggest that schools, where the NGO has delivered at least one flip chart, have higher scores than those without flip charts, and these results are statistically significant at 5%. Secondly, considering that our dependent variable is test scores normalized with respect to their mean¹, we have that ATE and ATT are estimated around ≈ 0.17 and 0.08 standard deviations, respectively. Most importantly, results are quite stable regardless of the number of nearest neighbors we use.

¹You can find a histogram for the distribution of test scores in Appendix I to verify they are centered to zero.

Table 2: MATCHING ESTIMATION

	Average Treatment Effect				Average Treatment on Treated			
	(1) 1 NN	(2) 3 NN	(3) 6 NN	(4) 10 NN	(5) 1 NN	(6) 3 NN	(7) 6 NN	(8) 10 NN
$\mathbb{1}(\text{Flip chart} = 1)$	0.166*** (0.012)	0.166*** (0.012)	0.169*** (0.012)	0.169*** (0.012)	0.085*** (0.015)	0.085*** (0.015)	0.086*** (0.015)	0.086*** (0.015)
N	52,820	52,820	52,820	52,820	52,820	52,820	52,820	52,820

Note: This table reports treatment effects using NN matching. The dependent variable is the normalized score. Each column represents an independent model that was run using 1, 3, 6, and 10 NN, respectively. Matching in characteristics is made using teacher training level, a dummy whether the classroom is indoor, a dummy indicating if the roof does not leak, the number of desks per student, the number of books per student, and the class size.

One relevant consideration when using matching methods is that, as with any other causal estimation tool, we need some “identifying assumptions”. In this case, the matching estimator relies on the so-called “ignorability” assumption which includes the unconfoundness (also known as conditional independence) and overlap (also known as common support) assumptions². With this in mind, remember that we want to be the closest possible to a random experiment to get an unbiased treatment effect. Hence, if the ignorability assumption holds, then after the matching process, we should expect similar covariates between our treatment units and the counterfactual control arm, as this would approximate us to a random experiment. Perfect balance is ensured by matching the characteristics of interest and then removing any potential selection bias. If this is true we expect a perfect balance post-matching, meaning that differences in means are close to zero and variance ratios close to one.

As we saw in *ayudantía 6* we can run this balancing test with the command `tebalance` in Stata. I implemented the post-matching balance test using the last model for the ATT with 10 neighbors (results are robust regardless of the number of neighbors). Results are reported in [Table 3](#). The first two columns display standardized differences for pre (raw) and post-matching for every covariate. We can see that before matching differences are large but after matching these immediately get close to zero. In addition, the variance ratio also improves after matching and gets close to one for almost every covariate.

Table 3: BALANCE POST-MATCHING

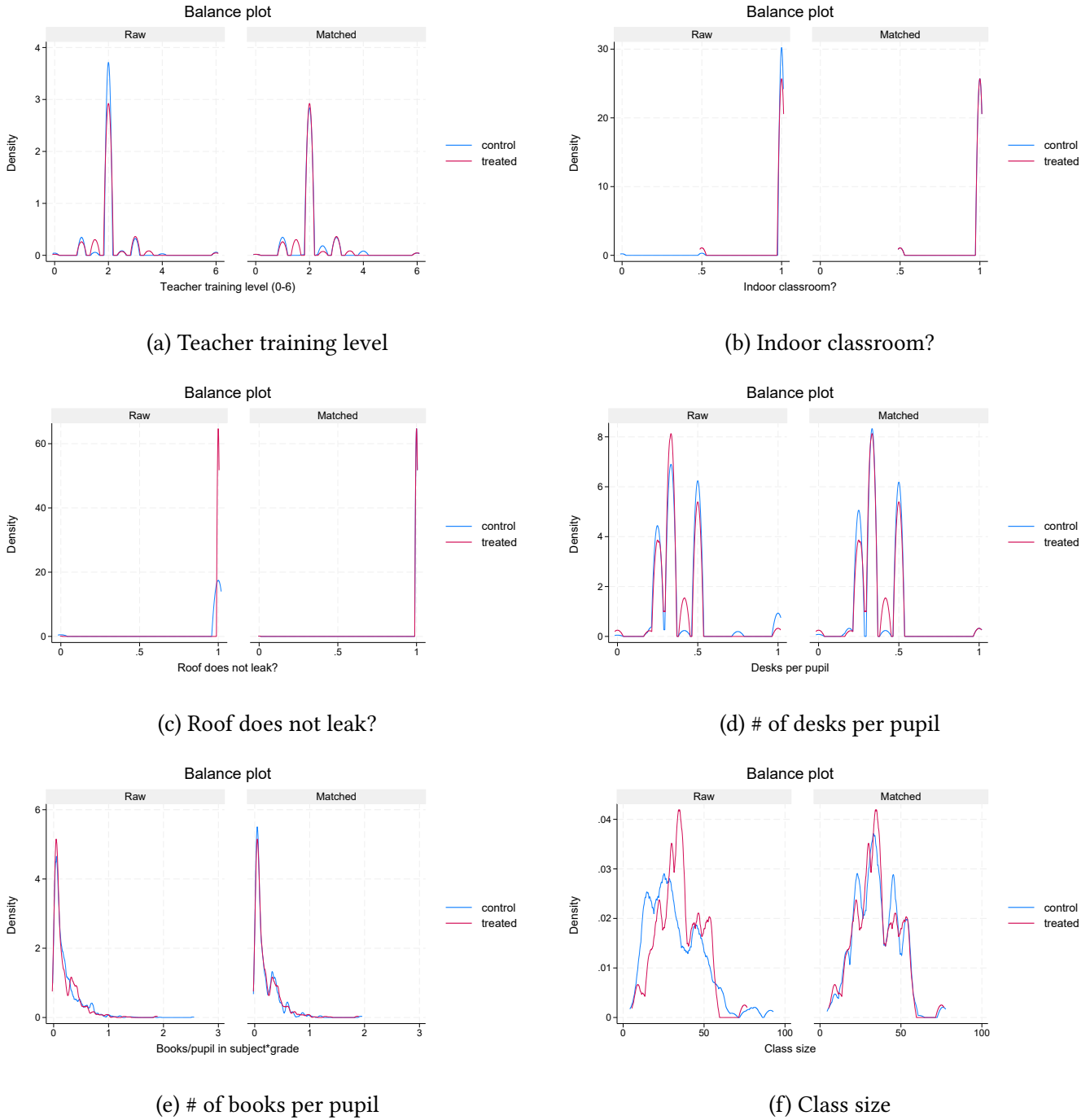
Covariate	Standardized differences		Variance ratio	
	Raw	Matched	Raw	Matched
Teacher training level	0.030	-0.054	1.005	0.899
Indoor classroom?	-0.081	0.000	1.013	1.000
Roof does not leak?	0.224	0.000	0.059	1.000
Desk per pupil	-0.205	-0.002	0.551	0.982
Books per pupil	-0.047	0.033	0.883	1.050
Class size	0.176	-0.007	0.606	0.988

Note: This table reports post-matching balance after implementing nearest-neighbors matching for 10 neighbors.

²Unconfoundness or Conditional Independence Assumption (CIA) is also known as the selection on observables assumption. This assumption states that, given a set of observable covariates (X), the treatment (T) is independent of the potential outcomes (Y). This can be expressed as $(Y_1, Y_0) \perp (T \mid X)$. On the other hand, the overlap assumption implies that for each value of the covariates (X), there is a probability related to receiving the treatment or not. Formally, this can be expressed as $0 < P(T = 1 \mid X) < 1$

The panel of Figure 1 shows graphic semi-parametric evidence for the balance pre- and post-matching for every covariate. Again, distributions after matching fit well between treatment and control groups. We could claim that our matching method reasonably corrects for the presence of selection bias, and the results discussed above in Table 2 are somewhat of a reasonable estimation of the treatment effects of receiving a flip chart. However, it is important to highlight that the results using matching are very sensitive to the metric distance we use, so we must be careful when trying to claim whether these are conclusive evidence for the treatment effects of the flip charts program.

Figure 1: DENSITIES FOR PRE- & POST-MATCHING



Note: This panel presents semi-parametric evidence for pre- and post-matching for every covariate included in the analysis.

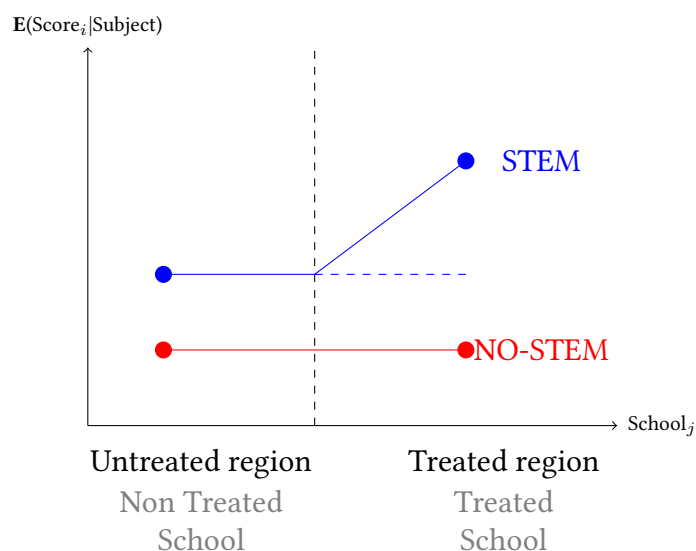
1.3 ALTERNATIVE EMPIRICAL STRATEGY

Usted se entera que los flip charts solo incluyen temas de matemáticas (mat), ciencias (sca) y tecnología (hsb) y no temas como inglés (eng), swahili (kis), geografía (ghc) ni artes (acm). ¿Cómo podría usar esta información para medir el impacto de los flip charts de una manera distinta a la realizada en a) y b)? Estime el impacto del programa usando esta metodología y explique los supuestos que deben cumplirse para lograr una correcta identificación. Realice un chequeo de robustez para convencer al lector de su investigación que su estrategia es válida.

The idea that comes to mind is to do a sort of Difference-in-Differences (DiD) based on [Lee & Kang \(2006\)](#). The intuition is as follows: to do a DiD, we need two dimensions. The first dimension relates to a “treatment region”. In the usual DiD for panel data, this treatment region is time --a region before the intervention and a region after the intervention-. The second dimension has to do with the expected value of an outcome that changes according to the treatment region it is in. This outcome for the second dimension tends to be at the individual level. The challenge is that we need to be able to construct its counterfactual if we want to obtain an estimate of the ATT.

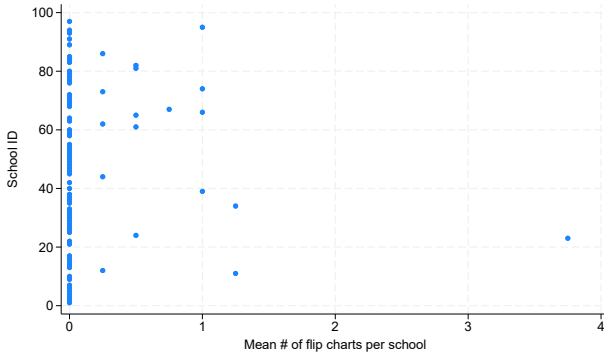
In this example case, we have the “second dimension”, which could be measured at the student, grade, or even at the subject level. Nonetheless, we do not have a panel dataset, thus we are missing a time dimension. Instead, we have an individual cross-section, so we need another definition for the “treatment region” which in this case will not be the time. For this purpose, I believe that we can use the treated and untreated schools as the treatment regions. This will allow me to disentangle between two discrete treatment regions more precisely, which is analogous to the case when we have pre- and post-treatment periods. The DiD I have in mind would look something like [Figure 2](#):

Figure 2: THEORETICAL ILLUSTRATION FOR DiD



Notice that the idea here is that in the absence of the flip charts, test scores of subjects related to science, technology, and math (from now on “STEM”) where the flip charts are included would have behaved the same as the test scores from subjects where flip charts are not included such as English, Swahili, geography, and arts (from now on “NO-STEM”). This argument is very similar to a “parallel trend assumption”, but ¿is it true that we can claim parallel trends in our example case? Or more generally speaking, ¿is it that the traditional identification assumptions for a DiD apply to our context? [Table 4](#) compares identification assumptions in the canonical model against the identification assumptions of my example case.

Table 4: Comparing Identification Assumptions

Canonical Assumptions	Our Example Case
<p>SUTVA: This assumption establishes that the treatment effect on a particular unit does not depend on the treatments received by other units in the sample. In other words, SUTVA assumes that there is no interference or external effects between treated and untreated units.</p>	<p>The data suggests that if a school appears with at least one flip chart then all classrooms in that same school have at least one flip chart (see figure down below). This could be interpreted as suggestive evidence that the treatment was assigned at the cluster (school) level. Thus, there are no plausible concerns of “spillover effects” within the school as all of the classrooms of treated schools have at least one flip chart.</p> 
<p>Overlap: We need each individual to have a non-negative probability of receiving the treatment.</p>	<p>This assumption is quite plausible enough because our treatment units (recall, the “second treatment dimension”), received the treatment according to very specific traits, namely whether it comes to a STEM subject or not.</p>
<p>Conditional Independence Assumption: It is assumed that, after controlling for certain covariates, the treatment is independent of the potential outcome.</p>	<p>We need to make sure that there are not any unmeasured confounders affecting our potential outcomes. We know that the treatment assignment is strongly related to specific observables of schools, so is reasonable to argue that the program is targeted to “poor/bad performing” schools. Hence, there might be other variables that are intrinsically related to these types of schools such as bad blackboard conditions, lower streams, etc. We need to control these outcomes to make the CIA plausible.</p>
<p>Parallel Trends: It is assumed that, in the absence of the event, the treated and untreated units would have followed parallel trajectories in their outcomes.</p>	<p>This assumption may be the most difficult to claim in our example case because it is closely related to a temporal dimension, and we are not working with panel data. More generally speaking, we are not able to claim a strong parallel trend assumption considering that the untreated region is basically “poor/bad performing” schools that maybe would behave worse anyway. Even though we are not positioned to offer strong empirical evidence for this identification assumption as we are working with cross-sectional data, it is reasonable to assume that, in the absence of the assignation of flip charts, scores from STEM subjects would behave similarly, on average, to NO-STEM subjects in treated and non-treated schools.</p>

I will estimate the causal effect of receiving any flip chart by running the following DiD from a two-way fixed effect (TWFE) model:

$$\text{test scores}_{ij} = \lambda_i + \delta_j + \gamma \cdot \{\text{stem} \times \text{any_flip}\}_{ij} + \Gamma\beta + \varepsilon_{ij} \quad (1)$$

Where test scores_{ij} are test scores for subject “ i ” in school “ j ”; λ_i is a fixed effect at the subject level - analogous to the fixed effect at the individual level; δ_j is a fixed effect at the school level; stem_i is a dummy variable equal to 1 if the subject is science, technology, or math, zero otherwise; any_flip_j is equal to 1 if the school is treated and zero otherwise; Γ is a vector of observable covariates at the classroom/school level (the same we use for the matching estimation); ε_{ij} are errors. Standard errors are clustered at the subject level. I interpret γ as the effect of the flip charts on test scores. Notice that our identification implies controlling for time-invariant unobserved characteristics associated with every subject and school (i.e. the fact that STEM courses tend to be harder in their contents when compared to NON-STEM courses, or in some schools, teachers are more capable to understand and better explain to students hard contents). [Lee & Kang \(2006\)](#) offers an analytical framework for estimating DiD with cross-sectional data along with further explanations on the assumptions that allow us to interpret γ as a causal effect.

Results are displayed in [Table 5](#). For comparison purposes, I report several specifications. In column (1) I report a standard OLS model that does not include any control nor fixed effects; column (2) is the same specification but includes controls. In column (3) I include fixed effects, but without controls. Finally, column (4) displays the main DiD model, as in [Equation 1](#). We can notice that our main specification attenuates the bias from standard OLS specification in column (2). I interpret this as suggestive evidence that after controlling for unobserved confounders we are isolating the true effect of the program captured by $\gamma \cdot \{\text{stem} \times \text{any_flip}\}_{ij}$. Hence, test scores in STEM subjects in treated schools are, on average, higher by ≈ 0.06 s.d. with respect to non-treated schools. But this result is noisily estimated as it only turns out to be statistically significant at the 10%.

Table 5: CROSS SECTION DIFFERENCE-IN-DIFFERENCES

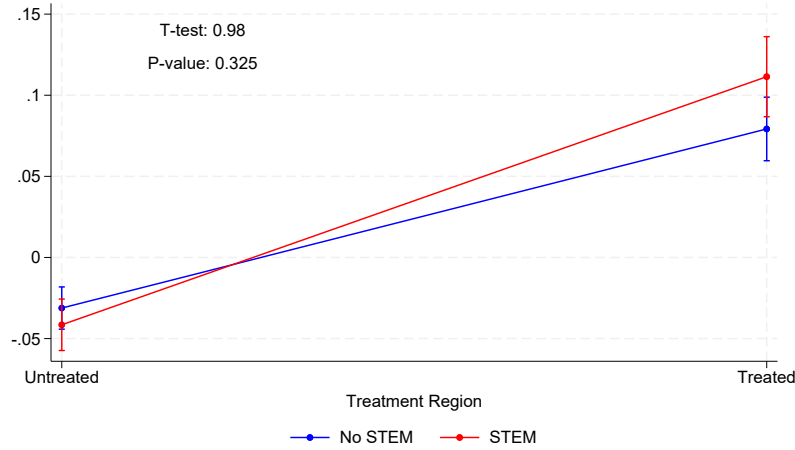
	(1)	(2)	(3)	(4)
$\text{stem} \times \text{any_flip}$	0.043 (0.029)	0.062* (0.029)	0.047 (0.026)	0.057* (0.025)
Observations	54,782	52,820	54,782	52,820
R-Squared	0.00	0.02	0.12	0.13
Subject FE?	No	No	Yes	Yes
School FE?	No	No	Yes	Yes
Controls?	No	Yes	No	Yes

Note: This table reports alternative specifications of [Equation 1](#).

Robustness Check. As I previously argued in [Table 4](#), we are not positioned to claim for the parallel trends assumption directly because of the absence of a time panel data dimension. However, we can rely on a loose intuition that is quite similar to the identification assumption of parallel trends. Since we have treated and untreated regions, and we are using NO-STEM subjects as the counterfactual, it is reasonable to assume that, in the absence of the assignation of flip charts, scores from STEM subjects would behave similarly, on average, to NO-STEM subjects in treated and non-treated schools. If this is true, then we would expect to see that both types of subjects (STEM or not) are not statistically different from zero in the untreated region, but different from zero in the treated region. This is an empirical question indeed. [Figure 3](#) reports graphic and statistical evidence for this assumption. A difference in means test suggests

that for coefficients in the untreated region, STEM and No STEM subjects are not statistically different from zero. Moreover, we can see that in the treated region, the conditional means for treated and untreated subjects diverge. In sum, we are not positioned to support a parallel trends assumption, but what I claim here is that I rely on this as suggestive evidence that my empirical strategy is a reasonable approach to estimate the causal effect of the flip charts on test scores once I am able to control for observable and non-observable confounders at the subject level as I try to do in my TWFE main specification.

Figure 3: CONDITIONAL MEANS



Note: This figure displays means for test scores conditional on treated and untreated arms. T-test and p-values for a difference in means test in the untreated region are also displayed.

1.4 ATE ESTIMATION

Tiempo después, la ONG aleatoriza la entrega de flip charts usando una lotería básica. Los resultados de esta se encuentran en la base `prosp-data-kremer.dta`. Estime el efecto promedio de tratamiento (ATE) de los flip charts sobre el puntaje en pruebas estandarizadas. Estudie además los efectos del programa para cada materia por separado. ¿Qué espera obtener? ¿Qué encuentra? ¿Hay algo que llame su atención? ¿Algo que le preocupe?

We know that when we conduct a randomization we have removed the selection bias as long as the implementation was correctly made and we are sure we had perfect compliance. Let us consider that we meet these criteria in our example case. Then, we can properly estimate an Average Treatment Effect (ATE), which is defined by:

$$\hat{ATE} = \mathbb{E}(y_i^T - y_i^{NT})$$

Notice the expression above corresponds to a difference in means between treated and control arms. In our case, every component of the expression is observed. We could get an empirical estimation of the ATE as follows:

Let us consider the linear model

$$y_i = \alpha + \beta T_i + \varepsilon_i \quad (2)$$

with the observational rule

$$y_i = T_i \cdot y_i(1) + (1 - T_i) \cdot y_i(0)$$

Then, assuming $\mathbb{E}[T_i' \varepsilon_i] = 0$ –which is plausible as the treatment was randomly assigned–, we arrive at

the expression for the ATE

$$\begin{aligned}
\hat{ATE} &= \mathbb{E}(y_i^T) - \mathbb{E}(y_i^{NT}) \\
&= \mathbb{E}(y_i|T_i = 1) - \mathbb{E}(y_i|T_i = 0) \\
&= \mathbb{E}[y_i(1)] - \mathbb{E}[y_i(0)] \\
&= \mathbb{E}[\alpha + \beta] - \mathbb{E}[\alpha] \\
\therefore \hat{ATE} &= \beta
\end{aligned}$$

Hence, I estimate ATE running by OLS a difference in means test under regression format as in [Equation 2](#). My prior is that $\hat{\beta} > 0$; meaning that I expect flip charts to have a positive effect on students' test scores. Results are reported in [Table 6](#). ATE estimates suggest that flip charts have zero effects on test scores as the β coefficient estimated for $\mathbb{1}(\text{Flip chart} = 1)$ turns out to be very close to zero (≈ 0.004 standard deviations) and statistically insignificant at any standard level. The result holds whether we estimate the difference model using robust –column (1)– or clustered standard errors at the school level –column (2)–. Again, I considered both specifications because the exercise is not clear enough about the way randomization was implemented, but I argue that they might have been randomly assigned at the school level since the data shows overlapping in the sense that all classrooms in the same school are completely treated or not.

Table 6: ATE ESTIMATES

	(1)	(2)
$\mathbb{1}(\text{Flip chart} = 1)$	-0.004 (0.005)	-0.004 (0.075)
Constant	0.000 (0.003)	0.000 (0.050)
N	195,204	195,204

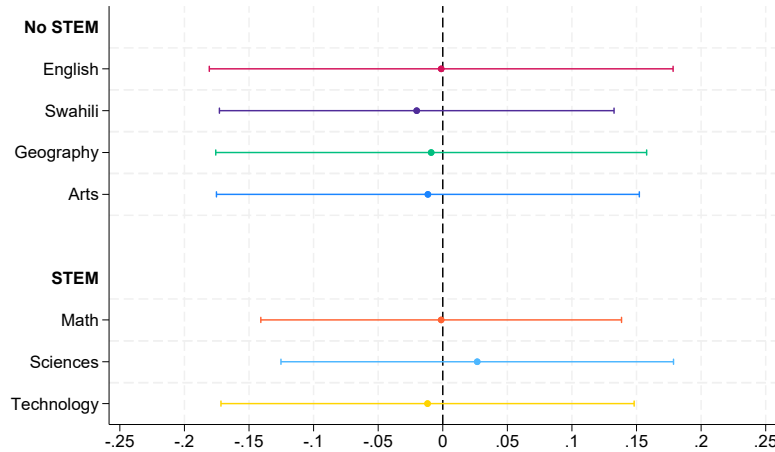
Note: This table reports ATE estimates as in [Equation 2](#). Column (1) includes robust standard errors, while column (1) includes clustered standard errors at the school level. No controls are included.

Heterogeneous effects by subject. Now, I run the [Equation 2](#) by each subject. In this case, I only run the specification clustering standard errors at the school level. [Figure 4](#) reports point estimates for ATE for independent regressions by subject along with 95% confidence intervals. Results are noisy and point estimates are very close to zero in magnitude. Again, we can not reject the null hypothesis of zero effects across subjects at any standard level of significance.

Concerns from Estimations. Opposite to my prior, flip charts do not significantly affect test scores. A priori, I would have expected null effects for NO-STEM subjects which are not covered in the contents of the flip charts. However, what is surprising to me is that results are not statistically significant even for any of the STEM classes such as sciences, technology, and math, which are the subjects which flip charts are supposed to be designed for. My concerns derived from these results are twofold. On one hand, from a methodological perspective, since we do not have any additional information regarding the design and implementation of the randomization I am concerned whether this is a plausible problem affecting our results. My intuition is that we could improve the precision of estimations by controlling for potential confounders or including strata fixed-effects, considering as strata the cluster at which the lotteries for assigning the flip charts were done, but the exercise does not provide further information on this. What we are doing here is leveraging randomness from a lottery assignation which is subject to imprecisions. For instance, we have 109k treated units and 119k untreated units, which might suggest the potential presence of unbalances that are related to a lack of power or minimum detectable effects estimations

before the assignment to determine the number of units in the treated and untreated groups. On the other hand, from an impact evaluation point of view, it seems like pushing for the flip chart as an educational program is not an effective tool to improve student performance. However, this conclusion must be taken carefully because it is hard for me to claim this is a true policy implication as I am not that sure whether we are facing a sharp RCT design.

Figure 4: TREATMENT EFFECTS BY SUBJECT



Note: This figure displays point estimates from independent regression of Equation 2 by subject. Standard errors are clustered at the school level.

1.5 COMPARISSON BETWEEN ESTIMATORS

Compare los estimadores del impacto obtenidos en las preguntas a), b), c) y d). ¿Qué puede decir respecto al sesgo de selección en este programa?

Table 7 displays the different estimates of treatment effects. Even though they are not directly comparable as each of the estimations relies on different identification assumptions, we can realize that results vary significantly quantitatively across methods. We are very tempted to believe that there is a positive effect of flip charts on test scores according to ATT estimations obtained from matching and DiD, but as I argued before, the matching results are very sensitive to the type of distance we use for the nearest neighbor, so it is not that obvious that we have robust evidence that matching is estimating a precise treatment effect. Moreover, the true impact estimate from the RCT indicates null effects.

In sum, this program has remarkable issues with selection bias as it tries to measure the impact of a tool that in most cases is related to other factors affecting student performance. I think that the easiest way to see the problems in terms of selection bias is by looking at the OLS results and comparing them with the results from the RCT. OLS is used as a first approximation to estimate the effect of flip charts on test scores without considering the endogeneity problem. Compared to ATE estimates using the RCT –which can be considered the “best” estimator of the treatment effect–, we can realize that upward bias accounts for a large fraction of the true average treatment effect.

Table 7: COMPARISSON BETWEEN ESTIMATES

OLS	Matching (ATT)	DiD (ATT)	RCT (ATE)
0.127	0.086***	0.057*	-0.004
(0.126)	(0.015)	(0.025)	(0.075)

2 INSTRUMENTAL VARIABLES

In this section we are asked to study IV methods using as a reference the paper “*The Colonial Origins of Comparative Development: An Empirical Investigation*” by [Acemoglu et al. \(2001\)](#).

2.1 OLS REGRESSIONS

Replique la tabla 2 (Columnas 1-6) del paper y discuta (siguiendo los argumentos de los autores) sobre la interpretación de los resultados en términos de causalidad.

The authors run regressions start by running OLS regressions to address possible endogeneity and selection bias issues in the analysis. The model takes the following form:

$$\log y_i = \mu + \alpha R_i + \mathbf{X}_i' \gamma + \varepsilon_i \quad (3)$$

where y_i is the income per capita in country i , R_i is the protection against expropriation measure, \mathbf{X}_i is a vector of other covariates, and ε_i is a random error term. The coefficient of interest throughout the paper is α , the effect of institutions on income per capita.

Requested replication is displayed in [Table 8](#). As suggested by the exercise we are not able to replicate the results reported by precisely [Acemoglu et al. \(2001\)](#), but they are somewhat close to those in the paper not only quantitatively but qualitatively as well. We observe a strong and statistically significant (at any standard level) correlation between institutions and economic growth of ≈ 0.4 - 0.5 log points. Nonetheless, we are aware that selection bias issues affect these results, so we can not claim that we are facing the causal effect of institutions on economic performance. Some of the reasons that make us believe that these are not a causal estimation are the following:

- Rich economies may be able to afford, or perhaps prefer, better institutions. This implies a reverse causality problem as better economic performance allows for investment in better institutions, but also strong institutions enhance democracy, and thus political and justice environment overall yields a well-functioning economic system.
- The authors also argue that the measures of institutions are constructed ex-post, and the analysts may have had a natural bias in seeing better institutions in richer places. In my interpretation, with this argument, the authors might suggest that there are potential measurement errors derived from the variables used in their estimations, and this can bias the parameter of interest.
- Their measure of institution is a cluster of a wide set of political, justice, and economic-related institutions so their index has a mixture of institutions that might not be that relevant for predicting economic growth in the long run. This also adds attenuation bias to the OLS estimations.

Table 8: OLS REGRESSIONS

	(1) Whole World	(2) Base Sample	(3) Whole World	(4) Whole World	(5) Base Sample	(6) Base Sample
Average protection against expropriation risk, 1985-1995	0.53*** (0.03)	0.52*** (0.05)	0.46*** (0.05)	0.41*** (0.05)	0.47*** (0.06)	0.40*** (0.06)
Latitude			0.87* (0.50)	0.54 (0.51)	1.58** (0.65)	0.88 (0.61)
Asia dummy				-0.24 (0.18)		-0.58* (0.30)
Africa dummy				-0.96*** (0.15)		-0.88*** (0.16)
Other continent dummy				-0.22 (0.20)		0.11 (0.22)
R^2	0.61	0.54	0.62	0.72	0.57	0.71
N	111	64	111	111	64	64

Note: This table replicates table 2 as in the paper of [Acemoglu et al. \(2001\)](#). In all cases, the dependent variable is the log GDP per capita (PPP basis) in 1995, current prices.

2.2 DISCUSSING THE INSTRUMENTS

Discuta brevemente el(los) instrumento(s) utilizado(s) por los autores, especialmente en relación a las condiciones que este(os) debe(n) cumplir para identificar insesgadamente el parámetro de interés.

The authors are interested in studying the effect of institutions on per capita income nowadays, but as argued above, the relationship in [Equation 3](#) is biased. In order to estimate an unbiased causal effect, [Acemoglu et al. \(2001\)](#) implement an Two-Stage Least Squares (2SLS) for:

$$\begin{aligned}\log y_i &= \mu + \alpha R_i + \mathbf{X}_i' \gamma + \varepsilon_i \\ R_i &= \zeta + \beta \log M_i + \mathbf{X}_i' \delta + v_i\end{aligned}\tag{4}$$

Where R is the measure of current institutions (protection against expropriation between 1985 and 1995). The authors suggest that the log of settler mortality ($\log M_i$) during the time of colonization is a plausible instrument for the measure of current institutions (R_i). This identification strategy will be valid as long as mortality rates faced by settlers are uncorrelated with the errors in [Equation 3](#). This is the first identification assumption $\mathbb{E}(R_i' M_i) \neq 0$, that mortality rates of settlers between the seventeenth and nineteenth centuries do not affect income today other than through their influence on institutional development. The intuition is that Europeans did not have an effective control over tropical diseases such as malaria or yellow fever, which accounted for almost 80% of settler deaths, but these diseases did not affect much to the natives. They argue that the mortality of local people and population densities before the arrival of Europeans are reasons to believe that settler mortality is a plausible instrument for institutional development: these diseases affected European settlement patterns and the type of institutions they set up, but had little effect on the health and economy of indigenous people.

The second identification assumption is an exclusion restriction $\mathbb{E}(M_i' \varepsilon_i) = 0$. Honestly, it was not very clear to me whether the authors addressed this assumption. The only thing I recall is that they state “the exclusion restriction is that this variable does not appear in [Equation 3](#)”. The intuition is that mortality rates of settlers in the 1800s are not correlated to any unobservable country characteristic nowadays.

2.3 2SLS vs OLS

Replique la tabla 4 (Columnas 1-8) del paper e interprete los resultados obtenidos con 2SLS vs OLS.

In this section, I estimate the 2SLS model as in [Equation 3](#) and [Equation 4](#). This replicates table 4 in the paper of [Acemoglu et al. \(2001\)](#). Results of my replication are displayed in [Table 9](#). Panel A of Table 4 reports 2SLS estimates of the coefficient of interest, α from [Equation 3](#). Panel B gives the corresponding first stages. Panel C displays OLS estimations. Column (1) displays the strong first-stage relationship between $\log M_i$ and current institutions.

The coefficients from the 2SLS indicate that the causal effect of institutions on economic development is 0.94 and significant at any standard level. Notice that this result is substantially larger than OLS estimates. reported in Panel C but also in [Table 8](#). The authors state that these differences in magnitudes evidence how measurement errors in the institution variables that create attenuation bias are likely to be more important than the causality reverse problem. This result varies in magnitude across the eight specifications –including covariates such as latitude and continent dummies, or after removing Neo-Europes or African continent– but is qualitatively robust as the author estimates a positive and statistically significant effect of institutions on economic performance. Again, every coefficient from the 2SLS is larger in magnitude than the OLS estimation.

Table 9: IV REGRESSIONS OF LOG GDP PER CAPITA

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Base Sample	Base Sample	Base Sample without Neo-Europes	Base Sample without Neo-Europes	Base Sample without Africa	Base Sample without Africa	Base Sample without Africa	Base Sample with continent dummies
Panel A: Two-Stage Least Squares								
Average protection against expropriation risk, 1985-1995	0.94*** (0.16)	1.00*** (0.22)	1.28*** (0.36)	1.21*** (0.35)	0.58*** (0.10)	0.58*** (0.12)	0.98*** (0.30)	1.11** (0.46)
Latitude		-0.65 (1.34)		0.94 (1.46)		0.04 (0.84)		-1.18 (1.76)
Asia dummy							-0.92** (0.40)	-1.05* (0.52)
Africa dummy							-0.46 (0.36)	-0.44 (0.42)
Other continent dummy							-0.94 (0.85)	-0.99 (1.00)
Panel B: First Stage								
Log European settler mortality	-0.61*** (0.13)	-0.51*** (0.14)	-0.39*** (0.13)	-0.39*** (0.14)	-1.21*** (0.22)	-1.14*** (0.24)	-0.43** (0.17)	-0.34* (0.18)
Latitude		2.00 (1.34)		-0.11 (1.49)		0.99 (1.43)		2.01 (1.39)
Asia dummy							0.33 (0.50)	0.47 (0.50)
Africa dummy							-0.27 (0.41)	-0.26 (0.41)
Other continent dummy							1.24 (0.84)	1.06 (0.84)
R^2	0.27	0.30	0.13	0.13	0.47	0.47	0.30	0.33
Panel C: Ordinary Least Squares								
Average protection against expropriation risk, 1985-1995	0.52*** (0.06)	0.47*** (0.06)	0.49*** (0.08)	0.47*** (0.07)	0.48*** (0.07)	0.47*** (0.07)	0.42*** (0.06)	0.40*** (0.06)
N	64	64	60	60	37	37	64	64

Note: This table replicates the table 4 as in the paper of [Acemoglu et al. \(2001\)](#). In all cases, the dependent variable is the log GDP per capita (PPP basis) in 1995, current prices.

2.4 WEAK INSTRUMENTS TEST

Para saber si estamos o no ante la presencia de instrumentos débiles, testee usando el estadístico de Cragg y Donald junto con las tablas de Stock y Yogo con respecto al tamaño o size del test de Wald. ¿Cómo testean los autores la fortaleza del instrumento? Compare ambos resultados.

We will test the identification assumption for an unbiased estimation of the treatment effect, with null hypothesis $H_0 : \mathbb{E}(R_i' M_i) \neq 0$. That is, the mortality rates of settlers need to be correlated with the risk of expropriation –our proxy for institutions-. Under the presence of weak instruments, the IV estimator is biased in the same direction as the OLS estimator and can even be inconsistent. Moreover, under the presence of weak instruments, the tests have a wrong significance and confidence intervals have wrong coverage probabilities. Thus, we need to make sure this is not the case.

The authors do not carry out any formal test (like Cragg and Donald) for this assumption. They rather corroborate graphically (in Figure 3 of the paper) the relationship between settler mortality rates and the index of institutions, and also run OLS regressions for Equation 4 –reported in columns (9) and (10) of Panel A, Table 3, and Panel B of Table 4 of the paper-. They show that settler mortality alone explains 27 percent of the differences in institutions observed today.

Now, I turn to test for the presence of weak instruments by implementing the Cragg & Donald test. This is obtained as the minimum eigenvalue of a concentration matrix. The null hypothesis is that we are in the presence of weak instruments. Stock & Yogo (2005) show that if we only have one endogenous variable, the lower eigenvalue of the concentration matrix is equal to the F-Statistic of the first stage of excluded instruments. Table 10 and Table 11 display the minimum eigenvalue of the concentration matrix (from all specifications in Table 9) and Stock & Yogo critical values, respectively. The results of Table 11 suggest that if we have one endogenous variable (as in our example case for R_i) along with one excluded instrument from the structural equation (in our case is M_i) and we want the Wald test with a nominal mean of 5% – on estimates for α in Equation 3– to restrict the bias of the 2SLS estimator to 5% of the OLS bias the critical value for the F-statistic in the first stage has to be ≈ 16.4 . Notice the F-statistics reported in Table 12 meet this requirement in three out of the eight specifications. Except for the last model for the base sample with continent dummies, all specifications meet an F-statistic which is higher than the 5.5 required to restrict the bias to 30% of the OLS estimator. In sum, we can conclude that we are not under the presence of weak instruments in the main model specification.

Table 10: CRAGG & DONALD TEST

	Base Sample	Base Sample	Base Sample without Neo-Europes	Base Sample without Neo-Europes	Base Sample without Africa	Base Sample without Africa	Base Sample without Africa	Base Sample with continent dummies
Min eigenvalue (F-stat)	22.95	13.09	8.65	7.83	30.54	21.61	6.23	3.46

Note: This table reports the minimum eigenvalue statistic from the Cragg & Donald concentration matrix.

Table 11: STOCK & YOGO TABLE

Critical values	5%	10%	20%	30%
2SLS size of nominal 5% Wald test	16.38	8.96	6.66	5.53
LIML size of nominal 5% Wald test	16.38	8.96	6.66	5.53

Note: This table reports critical values for minimum eigenvalue statistic as in Stock & Yogo (2005).

2.5 OVERIDENTIFICATION TEST

Replique la tabla 8 (excepto el panel C) del paper y comente sobre la interpretación que hacen los autores. Comente además acerca de los tests de sobreidentificación realizados aún cuando utilizan un solo instrumento.

My replication of Table 8 of the paper is here displayed in [Table 12](#). At this point, we need to remark that the authors made the overidentification tests for the following system:

$$\log y_i = \mu + \alpha R_i + \mathbf{X}_i' \gamma + \varepsilon_i$$

$$R_i = \lambda_R + \beta_R C_i + \mathbf{X}_i' \gamma_R + \nu_{Ri} \quad (5)$$

$$C_i = \lambda_C + \beta_C S_i + \mathbf{X}_i' \gamma_C + \nu_{Ci} \quad (6)$$

$$S_i = \lambda_S + \beta_S \log M_i + \mathbf{X}_i' \gamma_S + \nu_{Si} \quad (7)$$

Where we have settler mortality (M), affected settlements (S); settlements affected early institutions (C); and early institutions affected current institutions (R). Quoting the authors, the intuition they run after with the implementation of this test is the following:

“Overall, the overidentification test will reject the validity of our approach if either (i) the equation of interest, [Equation 3](#), does not have a constant coefficient, or (ii) C or S has a direct effect on income per capita, $\log y_i$ (i.e., either S or C is correlated with ε_i), or (iii) settler mortality, M , has an effect on $\log y_i$ that works through another variable, such as culture.”

What we are trying to test here is the exclusion restriction, with null hypothesis $H_0 : \mathbb{E}(M_i' \varepsilon_i) = 0$. Since there are a lot of regressions, let me focus on Panel D in [Table 12](#) which reports the results from 2SLS with European settler mortality rate as exogenous. This is an easy-to-interpret version of the overidentification test because if settlers' mortality rates had a direct effect on income per capita, this variable would be negative and significant. Notice that all coefficients turn out to be statistically insignificant. The author interprets these results as compelling evidence that the impact of mortality rates faced by settlers likely works through their effect on institutions and not by any other channel. That is, mortality rates are an exogenous variable and the exclusion restriction is met.

Over-identification tests with one instrument. Notice that in the paper of [Acemoglu et al. \(2001\)](#) they have up to three instrumental variables, but this is not the case in most of the empirical analyses we could make in real life. This is why we are asked to discuss the overidentification test even when we only have available one exogenous variable. In the context of IV estimation, overidentification tests, such as the Sargan or Hansen tests, are typically used when there are more instruments than endogenous variables. However, when there is only one instrument we would not be overidentified, and these tests are not applicable. The reason is that these tests are designed to assess the validity of the extra instruments by checking if the instruments as a group are uncorrelated with the error term. In a situation with a single instrument, the focus shifts to the relevance and validity of that instrument. The instrument should be strongly correlated with the endogenous regressor and uncorrelated with the error term. The validity of the instrument can be partially assessed using the first-stage F-statistic, which checks the strength of the instrument. A common rule of thumb is that the F-statistic should be greater than 10.

Table 12: OVERIDENTIFICATION TESTS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Two-Stage Least Squares										
Average protection against expropriation risk, 1985-1995	0.87*** (0.14)	0.92*** (0.20)	0.71*** (0.15)	0.68*** (0.20)	0.72*** (0.14)	0.69*** (0.19)	0.59*** (0.14)	0.61*** (0.17)	0.55*** (0.12)	0.55*** (0.14)
Latitude		-0.47 (1.24)		0.34 (1.08)		0.31 (1.05)		-0.40 (0.92)		-0.16 (0.81)
Panel B: First Stage										
European settlements in 1900	0.03*** (0.01)	0.03*** (0.01)								
Constraint on executive in 1900			0.32*** (0.08)	0.26*** (0.09)						
Democracy in 1900					0.24*** (0.06)	0.20*** (0.07)				
Constraint on executive in first year of indep.							0.25*** (0.08)	0.22*** (0.08)		
Democracy in first year of independence									0.19*** (0.05)	0.17*** (0.05)
R^2	0.30	0.30	0.20	0.24	0.24	0.26	0.19	0.25	0.26	0.30
Panel D: 2SLS with Log-Mortality as Exogenous										
Average protection against expropriation risk, 1985-1995	0.81*** (0.23)	0.88*** (0.29)	0.45* (0.25)	0.42 (0.30)	0.51** (0.23)	0.48* (0.27)	0.48** (0.23)	0.49* (0.25)	0.40** (0.18)	0.41** (0.19)
Log European settler mortality	-0.07 (0.16)	-0.05 (0.18)	-0.25 (0.16)	-0.26 (0.17)	-0.21 (0.15)	-0.22 (0.16)	-0.14 (0.16)	-0.14 (0.15)	-0.19 (0.13)	-0.19 (0.12)
Latitude		-0.52 (1.15)		0.38 (0.89)		0.28 (0.86)		-0.38 (0.84)		-0.17 (0.73)

Note: This table replicates the table 8 as in the paper of [Acemoglu et al. \(2001\)](#). In all cases, the dependent variable is the log GDP per capita (PPP basis) in 1995, current prices. All estimations are restricted to the base sample only. Notice that the Panel C of the table reported in the paper is excluded here, just as requested by the exercise.

2.6 HAUSMAN TEST

Usando los resultados de replicación de las Tablas 2 y 4, realice el test de Hausman e interprete sus resultados.

I run the Durbin-Wu-Hausman test for weak exogeneity, with null hypothesis $H_0 : \mathbb{E}(R_i' \varepsilon_i) = 0$, implying that the mortality rate is weakly exogenous and uncorrelated with the errors. If we reject the null hypothesis, then this indicates that there is endogeneity in the model. Results are displayed in Table 13. Column (1) of this table compares the model in column (2) of Table 8 vs column (1) in Table 9. Column (2) of this table compares the model in column (5) of Table 8 vs column (2) in Table 9. Column (3) of this table compares the model in column (6) of Table 8 vs column (8) in Table 9. We can see that in the first two specifications, the p-value is lower than 5%, thus we have evidence rejecting the null hypothesis of weak exogeneity, so the average protection against expropriation risk (R_i) is endogenous in those specifications. However, we are not able to reject H_0 for specification in column (3). Recall that this results from models including controls such as latitude, Asia dummy, Africa dummy, and other continent dummies. This result points out suggestive evidence that after including omitted variables (thus reducing uncertainty from unobservables) mortality rates could be an exogenous outcome, thus affecting some of the main identification assumptions of the paper.

Table 13: DURBIN-WU-HAUSMAN TEST

Statistic	(1)	(2)	(3)
χ^2	8.59	6.19	2.36
p-value	0.003	0.045	0.671

Note: This table reports critical values for Durbin-Wu-Hausman test.

2.7 COMMENTS FROM ALBOUY

Lea el paper The Colonial Origins of Comparative Development: An Empirical Investigation: Comment de David Albouy (2012). ¿Cuál es el argumento central del paper? ¿Qué hace el autor para cuestionar los resultados de Acemoglu et al. (2001)? ¿Qué puede decir respecto a sus resultados en d) luego de leer el paper?

Argument of the paper The central argument of Albouy (2012) paper is that the reliability and comparability of the European settler mortality rates used by Acemoglu et al. (2001) are questionable. He highlights several methodological issues, including the fact that many mortality rates were assigned to countries based on conjectures rather than actual data, increasing measurement errors. He also points out that the mortality rates used were not from actual European settlers but rather from soldiers and other groups, which may not accurately reflect the conditions experienced by settlers.

Questioning the paper of Acemoglu. Albouy (2012) examines settler mortality data and shows compelling evidence that the assignments of mortality rates to different countries made by Acemoglu et al. (2001) are often based on incorrect assumptions. He shows that when these questionable data points are removed or corrected, the strength of the relationship between settler mortality and expropriation risk diminishes significantly. Additionally, Albouy points out that controlling for the source of the mortality rates decreases the empirical relationship. In sum, he proves that results from Acemoglu et al. (2001) are not that robust.

What about our results above? Notice that in item d) of this problem we are asked to run a weak instrument test $\mathbb{E}(R_i' M_i)$ and we concluded that we were not under the presence of weak instruments.

According to [Albouy \(2012\)](#) arguments, is hard to claim that mortality rates are not weak instruments because as I mentioned in the paragraph above after [Albouy \(2012\)](#) controls for the source of the mortality rates further weakens the empirical relationship between mortality rates and expropriation risks.

2.8 RESULTS FROM ALBOUY

Utilizando la base de datos `albouy_2012.dta` replique la tabla 2 (Columnas 1-4, paneles A-D) de [Albouy \(2012\)](#) y comente sobre los resultados obtenidos.

Table 2 from [Albouy \(2012\)](#) is displayed in [Table 14](#). In Panel A we have the original dataset used by [Acemoglu et al. \(2001\)](#) for different specifications of the first stage using the expropriation risk as the dependent variable. Since there are a lot of regressions, let me be more general in my interpretation here. In the first place, we observe that point estimates are substantially over-estimated in the original data and sensitive to the inclusion or exclusion of some controls or subgroups. The variance is also sensitive and imprecisely estimated. For instance, if we drop conjectured mortality rates, the conjectured mortality rates appear to mask the collinearity between the controls and the more accurately measured rates. Using the original sample again, in Panel C [Albouy \(2012\)](#) demonstrates that controlling for whether a mortality rate comes from soldiers on campaign or from African laborers makes log mortality insignificant at the 5 percent level in all specifications with controls. Finally, results in Panel D suggest that the campaign and laborer indicators become much more significant once the conjectured data are dropped. In sum, the estimates reported by [Acemoglu et al. \(2001\)](#) are extremely sensitive to the data source and empirical specification. Again, after this evidence is hard to claim that results reported by [Acemoglu et al. \(2001\)](#) are robust but even more, is harder to claim whether the mortality rate is a good instrument overall.

Table 14: FIRST STAGE ESTIMATES

Control variables	No controls	Latitude control	Without Neo-Europes	Continent indicators
Panel A. Original data				
Log mortality (β)	-0.61	-0.51	-0.39	-0.43
homocedastic standard error	0.13	0.14	0.13	0.17
heterocedastic-clustered SE	0.17	0.19	0.17	0.20
p-value of log mortality	0.00	0.01	0.03	0.04
p-value of controls	-	0.18	-	0.40
Panel B. Removing conjectured mortality rates				
Log mortality (β)	-0.59	-0.42	-0.32	-0.31
heterocedastic standard error	0.19	0.22	0.19	0.20
p-value of log mortality	0.01	0.07	0.10	0.13
p-value of controls	-	0.05	-	0.01
Panel C. Original data, adding campaign and laborer indicators				
Log mortality (β)	-0.43	-0.37	-0.29	-0.35
heterocedastic clustered SE	0.18	0.20	0.17	0.21
p-value of log mortality	0.02	0.07	0.10	0.11
p-value of indicators	0.17	0.22	0.29	0.26
p-value of controls	-	0.25	-	0.73
Panel D. Removing conjectured mortality and adding campaign and laborer indicators				
Log mortality (β)	-0.35	-0.21	-0.18	-0.25
heterocedastic standard error	0.22	0.25	0.22	0.23
p-value of log mortality	0.12	0.42	0.42	0.28
p-value of indicators	0.03	0.06	0.08	0.34
p-value of controls	-	0.08	-	0.03

Note: This table replicates table 2 as in [Albouy \(2012\)](#).

3 REGRESSION DISCONTINUITY

In this section, we are asked to replicate the paper “*The effects of drinking and driving laws on car crashes, injuries, and deaths: Evidence from Chile*” by [Otero & Rau \(2017\)](#).

3.1 EMPIRICAL STRATEGY

Describe brevemente el objetivo del paper, la estrategia de identificación. Refiérase a la “running variable” y al diseño empírico.

The main goal of the paper is to analyze the causal effects of a Law in Chile that was enacted in 2012 that decreased the legal blood alcohol content limit for drivers from 0.05 to 0.03 grams of alcohol per deciliter of blood (g/dL) and increased license suspension periods for people committing these offenses. The authors exploit rich data on car accidents and their causes and combine it with three empirical methods such as regression discontinuity (RD), generalized Poisson, and linear logit regressions.

Among the three empirical methods the authors implemented, the one of our interest is RD. As a first estimation approach, the authors implement a sharp RD using the time (measured in weeks) as the running variable, exploiting the discontinuous change due to the law’s approval date (the fifth week of January 2012). This approach allows them to estimate the short-term causal effect of the law on several outcomes of interest such as accidents, injuries, and deaths.

3.2 IDENTIFICATION ASSUMPTIONS

Escriba la ecuación (1) del paper, explicando sus componentes, y demuestre que δ corresponde al efecto causal del programa en el contexto de una regresión discontinua. Sea formal y explícita en los supuestos necesarios para poder identificar δ . ¿Son plausibles?

The model to be estimated is as follows:

$$y_t = \alpha + \beta \cdot \text{Police}_{t-1} + \gamma \cdot \text{Gas}_t + \delta \cdot \text{Post}_t + g(t) + \epsilon_t \quad (8)$$

Where y_t is the number of accidents at time t ; $police_{t-1}$ is the number of police stops in time $t - 1$ (to avoid simultaneity issues), and gas_t is gasoline sales. The dummy variable $post_t$ is equal to one if the period t is after the law’s approval and equal to zero otherwise. Lastly, $g(t)$ is a smooth function of t . Some special attention deserves the component $g(t)$, which is a smooth function of the time. In this case, the authors implement Local Linear Regressions (LLR). Our parameter of interest is δ , which is aimed at capturing the causal effect of the program. Next, we are going to prove that δ estimates the causal effect of the Law. Let us assume that the following limits exist:

$$\lim_{t \rightarrow t_0^+} \mathbb{E}(\text{Post}_t \mid t_i = t) < \infty$$

$$\lim_{t \rightarrow t_0^-} \mathbb{E}(\text{Post}_t \mid t_i = t) < \infty$$

Where $t_0^+ \neq t_0^-$, and

$$\lim_{t \rightarrow t_0^+} \mathbb{E}(y_t \mid t_i = t) = \lim_{t \rightarrow t_0^+} \mathbb{E}[\alpha + \beta \cdot \text{Police}_{t-1} + \gamma \cdot \text{Gas}_t + \delta \cdot \text{Post}_t + g(t) + \epsilon_t \mid t_i = t] \quad (9)$$

$$\lim_{t \rightarrow t_0^-} \mathbb{E}(y_t \mid t_i = t) = \lim_{t \rightarrow t_0^-} \mathbb{E}[\alpha + \beta \cdot \text{Police}_{t-1} + \gamma \cdot \text{Gas}_t + \delta \cdot \text{Post}_t + g(t) + \epsilon_t \mid t_i = t] \quad (10)$$

Making $\mathbf{X}_t\Gamma \equiv \alpha + \beta \cdot \text{Police}_{t-1} + \gamma \cdot \text{Gas}_t$, we can simplify

$$\begin{aligned}\lim_{t \rightarrow t_0^+} \mathbb{E}(y_t | t_i = t) &= \lim_{t \rightarrow t_0^+} \mathbb{E}[\mathbf{X}_t\Gamma + \delta \cdot \text{Post}_t + g(t) + \epsilon_t | t_i = t] \\ \lim_{t \rightarrow t_0^-} \mathbb{E}(y_t | t_i = t) &= \lim_{t \rightarrow t_0^-} \mathbb{E}[\mathbf{X}_t\Gamma + \delta \cdot \text{Post}_t + g(t) + \epsilon_t | t_i = t]\end{aligned}$$

Subtracting Equation 9 from Equation 10:

$$\begin{aligned}\lim_{t \rightarrow t_0^+} \mathbb{E}(y_t | t_i = t) - \lim_{t \rightarrow t_0^-} \mathbb{E}(y_t | t_i = t) &= \lim_{t \rightarrow t_0^+} \mathbb{E}(\mathbf{X}_t\Gamma | t_i = t) - \lim_{t \rightarrow t_0^-} \mathbb{E}(\mathbf{X}_t\Gamma | t_i = t) \\ &\quad + \delta \left[\lim_{t \rightarrow t_0^+} \mathbb{E}(\text{Post}_t | t_i = t) - \lim_{t \rightarrow t_0^-} \mathbb{E}(\text{Post}_t | t_i = t) \right] \\ &\quad + \lim_{t \rightarrow t_0^+} \mathbb{E}(g(t) | t_i = t) - \lim_{t \rightarrow t_0^-} \mathbb{E}(g(t) | t_i = t) \\ &\quad + \lim_{t \rightarrow t_0^+} \mathbb{E}(\epsilon_t | t_i = t) - \lim_{t \rightarrow t_0^-} \mathbb{E}(\epsilon_t | t_i = t)\end{aligned}\tag{11}$$

Notice that in the expression above, we have assumed that Post_t and δ are independent, but also using the assumption that the functions $\mathbb{E}(\mathbf{X}_t\Gamma | t)$, $\mathbb{E}(g(t) | t)$, $\mathbb{E}(\epsilon_t | t)$ are continuous at $t = t_0$ we have

$$\begin{aligned}\lim_{t \rightarrow t_0^+} \mathbb{E}(\mathbf{X}_t\Gamma | t_i = t) &= \lim_{t \rightarrow t_0^-} \mathbb{E}(\mathbf{X}_t\Gamma | t_i = t) \\ + \lim_{t \rightarrow t_0^+} \mathbb{E}(g(t) | t_i = t) &= \lim_{t \rightarrow t_0^-} \mathbb{E}(g(t) | t_i = t) \\ \lim_{t \rightarrow t_0^+} \mathbb{E}(\epsilon_t | t_i = t) &= \lim_{t \rightarrow t_0^-} \mathbb{E}(\epsilon_t | t_i = t)\end{aligned}$$

Rearranging Equation 11, we arrive to an expression for δ :

$$\begin{aligned}\lim_{t \rightarrow t_0^+} \mathbb{E}(y_t | t_i = t) - \lim_{t \rightarrow t_0^-} \mathbb{E}(y_t | t_i = t) &= \delta \left[\lim_{t \rightarrow t_0^+} \mathbb{E}(\text{Post}_t | t_i = t) - \lim_{t \rightarrow t_0^-} \mathbb{E}(\text{Post}_t | t_i = t) \right] \\ \delta &= \frac{\lim_{t \rightarrow t_0^+} \mathbb{E}(y_t | t_i = t) - \lim_{t \rightarrow t_0^-} \mathbb{E}(y_t | t_i = t)}{\lim_{t \rightarrow t_0^+} \mathbb{E}(\text{Post}_t | t_i = t) - \lim_{t \rightarrow t_0^-} \mathbb{E}(\text{Post}_t | t_i = t)}\end{aligned}$$

In a sharp design, like in our example case, we need monotonicity to make sure we are precisely estimating a causal effect. This condition is assured as the treatment is the period after the Law's approval.

$$\text{Post}_t = \begin{cases} 0 & \text{if } t_i \leq t_0 \\ 1 & \text{if } t_i > t_0 \end{cases}$$

Which implies that

$$\begin{aligned}\lim_{t \rightarrow t_0^+} \mathbb{E}(\text{Post}_t | t_i = t) &= 1 \\ \lim_{t \rightarrow t_0^-} \mathbb{E}(\text{Post}_t | t_i = t) &= 0\end{aligned}$$

Finally, we have arrived at an expression for the ATE at the cutoff:

$$\therefore \delta = \lim_{t \rightarrow t_0^+} \mathbb{E}(y_t | t_i = t) - \lim_{t \rightarrow t_0^-} \mathbb{E}(y_t | t_i = t)$$

This proves that δ estimates the causal effect of the Law. Overall, I believe that the identification assumptions behind the RD method are quite plausible. For instance, in this context, we rely on the assumption

that the assignation of the treatment is based only on the switching regime due to the law's approval date. In other words, there is a discontinuous assignment to treatment probability in a variable known as the running or forcing variable. Another assumption is that the outcomes of interest are continuous around the Law's approval date and citizens can not manipulate the approval of the Law in the sense that they can not sort into treatment status and that drivers do not change their behavior before the law was approved.

3.3 LOCAL LINEAR REGRESSION

Los autores eligen el método de “regresión lineal local” para estimar el efecto causal del program. ¿Cuál es la intuición de esta estimación? ¿Cómo eligen el ancho de banda los autores y cuál es la importancia de esto? ¿Qué rol juega la elección del Kernel?

The intuition of LLR. Local linear regression in RD focuses on fitting linear models to subsets of data close to the cutoff point. The basic idea is that by concentrating the estimation on a small neighborhood around the cutoff, we can obtain a more accurate and less biased estimate of the causal effect. Recall that our causal effect is given by

$$\lim_{t \rightarrow t_0^+} \mathbb{E}(y_t | t_i = t) - \lim_{t \rightarrow t_0^-} \mathbb{E}(y_t | t_i = t) = \hat{a}_L - \hat{a}_R$$

where these values are estimated from optimization problems of the form:

$$\begin{aligned} (\hat{a}_L, \hat{b}_L) &= \arg \min_{a,b} \sum_{i=1}^n \mathbb{1}(t_i \leq t_0) \cdot (y_i - a - b(t_i - t_0))^2 \cdot K\left(\frac{t_i - t_0}{h}\right) \\ (\hat{a}_R, \hat{b}_R) &= \arg \min_{a,b} \sum_{i=1}^n \mathbb{1}(t_i > t_0) \cdot (y_i - a - b(t_i - t_0))^2 \cdot K\left(\frac{t_i - t_0}{h}\right) \end{aligned}$$

where $\mathbb{1}(\cdot)$ is an indicator function that takes the value of 1 according to the region of interest -before or after the cutoff- $K(\cdot)$ is a kernel function, and h is the size of the bandwidth.

Bandwidth (h). The authors choose the robust optimal bandwidth of [Calonico et al. \(2014\)](#) for their estimations. This method minimizes mean squared error. The bandwidth in RD determines the range of data around the cutoff used for local estimation. A smaller bandwidth reduces bias by focusing on observations very close to the cutoff but may increase variance due to the smaller number of observations.

Kernel (K). The kernel is a weighting function that determines how observations within the bandwidth are weighted. The most common are triangular, uniform, and Epanechnikov. In the baseline results of the paper, [Otero & Rau \(2017\)](#) implement a triangular Kernel on both sides of the discontinuity.

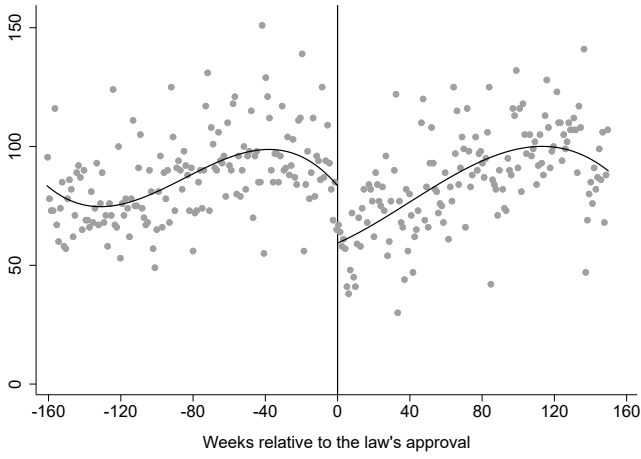
3.4 REPLICATING FIGURES

Replique la Figura 1 completa y el panel (a) de la Figura 2 del paper. Sea cuidadoso con el orden del polinomio de la running variable. ¿Qué puede decir de los efectos del programa en los outcomes de interés?

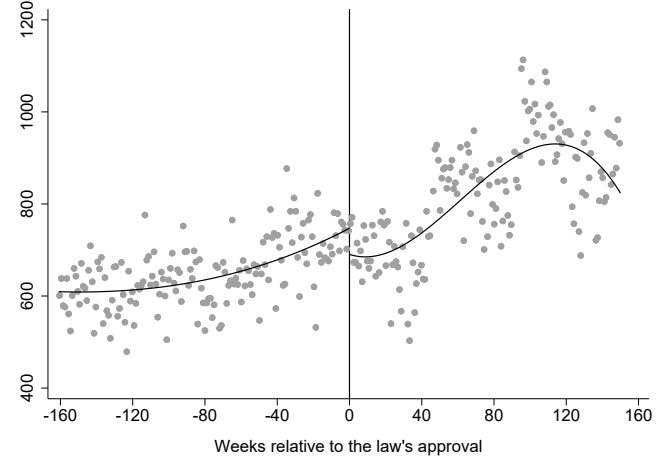
Panel of Figure 5 replicates requested illustrations. The graphics are for weekly data using a cubic specification for $g(t)$ and the law's approval as the cutoff date, as in Otero & Rau (2017). Each dot represents the number of car accidents in a given week. In the context of RD the graphical evidence is an important tool in order to (hopefully) offer compelling evidence of the potential effects of the treatment of our interest. In the figures, we can see three relevant elements: the cutoff of the running variable centered around zero indicating the Law's approval, the dispersion in the outcome, and, the fit of the LLR. Salient gaps around the cutoff would be suggestive evidence of the impact of the Law and vice-versa.

For the case of alcohol-related accidents -panel (a)- notice that the fitted line around the threshold after the implementation of the Law is located below the fitted line for the pre-reform period. This might suggest that the treatment region (post-law period) alcohol-related accidents decreased. On the other hand, panel (c) does not show any gap around the threshold so we should expect to see null effects around the cutoff. Finally, for all injuries -panel (e)-, again we observe a relevant gap around the cutoff suggesting a negative trend in all injuries post Law. Thus, we would expect to see a statistically significant effect of the Law for these outcomes in specific (accidents, alcohol-relate, and all injuries that are alcohol-related). We will formally test for these priors in the following question.

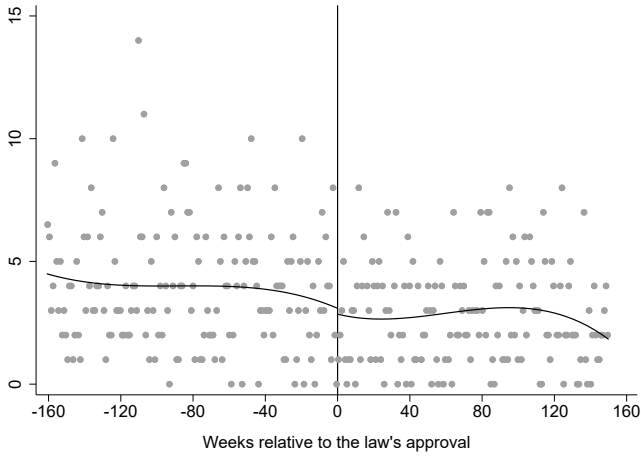
Figure 5: RD Plots



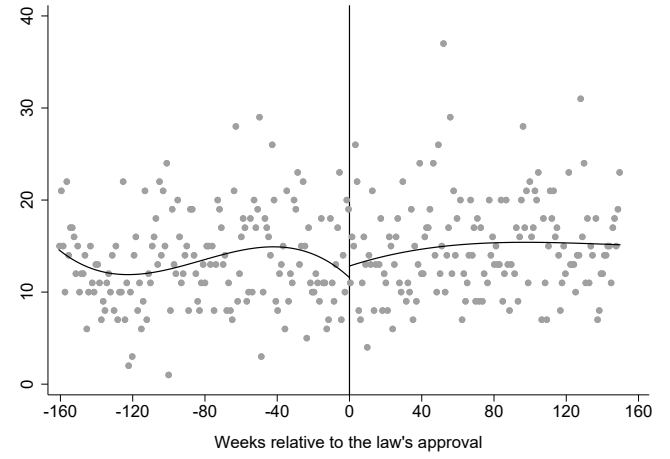
(a) Accidents, alcohol related



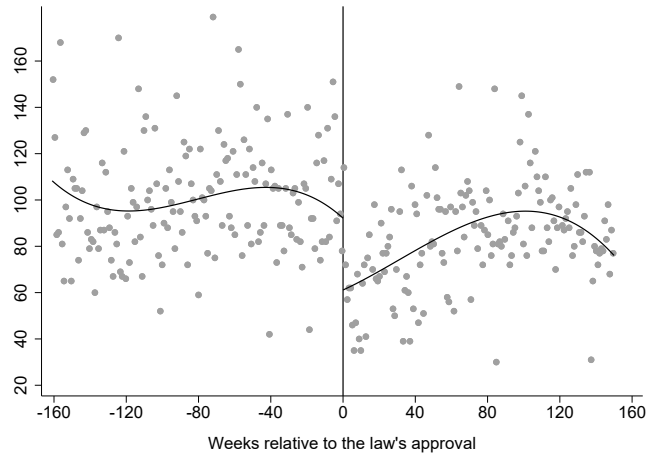
(b) Accidents, non-alcohol related



(c) Deaths, alcohol related



(d) Deaths, non-alcohol related



(e) All injuries, alcohol related

Note: This panel replicates figure 1 and 2a, as in [Otero & Rau \(2017\)](#).

3.5 REPLICATING RD RESULTS

Replique las columnas 1, 2 y 6 de la Tabla 5 del paper. ¿Cómo se interpretan los resultados?

Replication of the table of the paper is displayed in [Table 15](#). We can see the point estimates using LLR and the optimal bandwidth selection method by [Calonico et al. \(2014\)](#). It also includes the standard errors, the bandwidth, and the pre-law levels. Opposite to what the authors did for the paper, I estimate pre-law levels as the unconditional mean of the outcome between the first week of 2012 and the week right before the Law's approval, and they turn out the same. Point estimates from the RD are statistically significant at any standard level for alcohol-related accidents with a reduction of 26 accidents per week right after the law's approval, which is equivalent to $\approx 32\%$ decrease. For alcohol-related injuries, the authors also find a substantial and statistically significant decrease of ≈ 38 injuries right after the Law's approval, which is equivalent to $\approx 35\%$ decrease. However, there is not any significant effect of the law on deaths (at the 10% level).

It is important to take into account that these results can be understood as an Average Treatment Effect (ATE) around the cutoff, that is, a local treatment effect. And because our running variable is time, these are interpreted as short-term effects of the Law. Thus, the fact that we do not observe any immediate effect on alcohol-related deaths does not imply that there is an effect in the long run, for estimating the long-term effect we need another empirical strategy.

Table 15: REGRESSION DISCONTINUITY ESTIMATES

	Accidents	Injuries	Deaths
RD estimate	-26.06*** (5.56)	-37.92*** (10.49)	-0.48 (1.20)
Bandwidth	38.27	40.95	39.14
Pre-law level	82.25	107.00	3.75

Note: This table replicates table 5 as in [Otero & Rau \(2017\)](#).

3.6 CHANGING THE BANDWIDTH ESTIMATION

Estime los resultados presentados en la primera columna de la Tabla 5 manteniendo el ancho de banda utilizado, pero estimando con Kernel Uniforme, Triangular, y Epanechnikov. Comente sobre sus resultados.

Results are displayed in [Table 16](#). The first column replicates the baseline results as in [Table 15](#) using the triangular Kernel. The estimates from the three kernels are quite similar in magnitude, all indicating a reduction of around 26 to 27 alcohol-related accidents. The Epanechnikov kernel gives the highest reduction estimate (-27.31), followed closely by the Uniform (-26.83) and Triangular (-26.06) kernels. The standard errors are also similar, with the Uniform kernel having the highest standard error (6.11), followed by the Epanechnikov (5.78) and Triangular (5.56) kernels. The differences in standard errors are relatively small, indicating similar levels of precision in the estimates across the kernels. The consistency of the estimates across different kernels suggests that the estimated effect of the treatment on reducing alcohol-related accidents is somewhat robust. In any case, I think is important to remark that the choice of kernel impacts the estimates by balancing the trade-off between bias and variance differently. Triangular and Epanechnikov kernels are generally better for capturing local treatment effects with less bias. In contrast, the uniform kernel can provide more stable estimates but estimates a higher variance in this case.

Table 16: RD ESTIMATES - CHANGING KERNEL

	Triangular	Uniform	Epanechnikov
Alcohol related accidents	-26.06*** (5.56)	-26.83*** (6.11)	-27.31*** (5.78)
Bandwidth	38.27	38.27	38.27
Pre-law level	82.25	82.25	82.25

Note: This table replicates table 5 as in [Otero & Rau \(2017\)](#), but using alcohol-related accidents as the dependent variable and changing Kernels.

3.7 USING IK METHOD

Ahora estime los resultados de la primera columna de la Tabla 5, pero utilizando el ancho de banda conocido como IK. ¿Cambian sus resultados? ¿Cuál es la importancia de la elección del ancho de banda? ¿Cómo se diferencia éste método de calculo del ancho de banda con el utilizado en el paper?

Now I change baseline estimates for the number of alcohol-related accidents with different Kernels and also selecting the optimal bandwidth using [Imbens & Kalyanaraman \(2011\)](#) method. Results are displayed in [Table 17](#). Recall that the baseline point estimate was -26.06 with a SE of 5.56. Now, when implementing the IK procedure the magnitude of the point estimate increases up to 28 no matter the Kernel type. Also, standard errors are slightly higher.

Table 17: RD ESTIMATES - USING IK BANDWIDTH SELECTION

	Triangular	Uniform	Epanechnikov
Alcohol related accidents	-28.44*** (5.54)	-28.08*** (6.09)	-27.99*** (6.63)
Bandwidth	90.69	82.87	107.68
Pre-law level	82.25	82.25	82.25

Note: This table replicates table 5 as in [Otero & Rau \(2017\)](#), but using alcohol-related accidents as the dependent variable and changing Kernels.

¿Why the bandwidth is so important? In RD designs, the estimation of the Kernel is not as relevant as the bandwidth selection. This is because of the well-known trade-off between bias-variance in non-parametric models. A higher bandwidth includes a larger amount of observations in our estimation which improves the efficiency (lower variance) but increases bias as includes observations that are farther away from the cutoff adding units that are more different.

Differences between CCT and IK bandwidth selection methods. [Otero & Rau \(2017\)](#) estimate the causal effect of the driving Law in Chile by implementing the [Calonico et al. \(2014\)](#) method to select the bandwidth. The difference is that [Imbens & Kalyanaraman \(2011\)](#) derive an optimal bandwidth minimizing a version of the Mean Squared Error (MSE) but their pilot bandwidth is not optimal, so [Calonico et al. \(2014\)](#) improved their method that minimizing an asymptotic version of the MSE. To a better understanding of the differences between both methods, let me rely on [Calonico \(2014\)](#) who developed a command in Stata for RD analysis and address the differences between both methods:

Imbens & Kalyanaraman (2011) provide a data-driven, asymptotically MSE-optimal, RD treatment-effect estimator. Specifically, they propose a more "robust" consistent bandwidth estimator of the form

$$\hat{h}_{IK,n,p} = \left\{ \frac{\hat{V}_{IK,p}}{2(p+1)\hat{B}_{IK,p}^2 + \hat{R}_{IK,p}} \right\}^{\frac{1}{(2p+3)}} n^{\frac{-1}{(2p+3)}}$$

where the additional (regularization) term $\hat{R}_{IK,p}$ is introduced to avoid small denominators in moderate-size samples. Here $\hat{B}_{IK,p}$ and $\hat{V}_{IK,p}$ (and $\hat{R}_{IK,p}$) are nonparametric consistent estimators of their respective population counterparts, which require the choice of preliminary bandwidths, generically denoted by b_n herein. Imbens & Kalyanaraman (2011) provide a direct implementation approach for $p = 1$, but the preliminary bandwidths used in their construction are not optimally chosen. Thus $\hat{h}_{IK,n,p}$ can be viewed as a nonparametric first-generation plug-in rule, sometimes denoted by a direct plug-in rule of order 1.

Motivated by the work of Imbens & Kalyanaraman (2011), Calonico et al. (2014) propose an alternative second-generation plug-in bandwidth selection approach. Specifically, they propose the following second-order direct plug-in rule:

$$\hat{h}_{CCT,n,p} = \left\{ \frac{\hat{V}_{CCT,p}}{2(p+1)\hat{B}_{CCT,p}^2 + \hat{R}_{CCT,p}} \right\}^{\frac{1}{(2p+3)}} n^{\frac{-1}{(2p+3)}}$$

This alternative bandwidth estimator has two distinct features relative to $\hat{h}_{IK,n,p}$. First, not only are the estimators $\hat{V}_{CCT,p}$ and $\hat{B}_{CCT,p}$ (and $\hat{R}_{CCT,p}$) consistent for their population counterparts, but the preliminary bandwidths used in their constructions are consistent estimators of the corresponding population MSE-optimal bandwidths. In this sense, $\hat{h}_{CCT,n,p}$ is a direct plug-in rule of order 2.

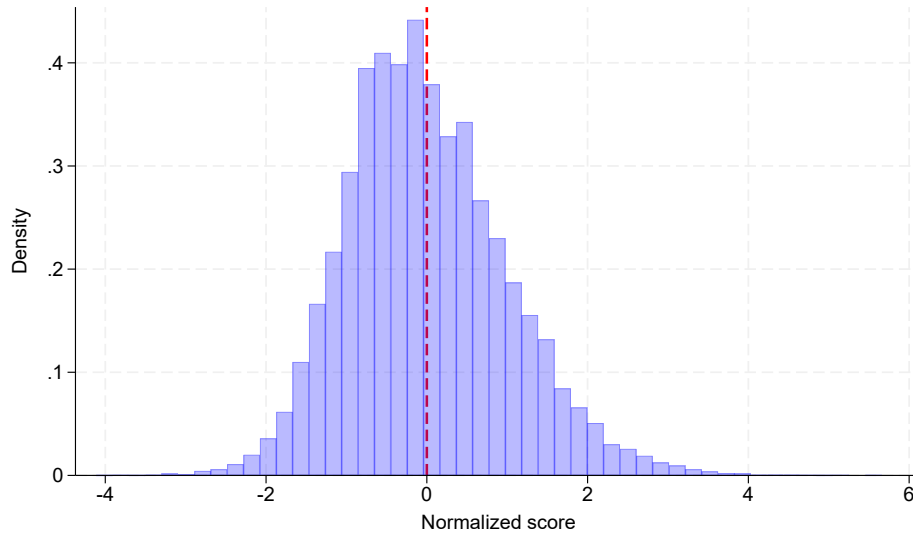
References

- Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1), 235–267. [3](#)
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5), 1369–1401. [12](#), [13](#), [14](#), [16](#), [17](#), [18](#), [19](#)
- Albouy, D. Y. (2012). The colonial origins of comparative development: An empirical investigation: Comment. *American Economic Review*, 102(6), 3059–76. [18](#), [19](#)
- Calonico, S. (2014). Robust data-driven inference in the regression-discontinuity design. *Stata Journal*, 14(4), 909–946(38). [26](#)
- Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6), 2295–2326. [22](#), [25](#), [26](#), [27](#)
- Imbens, G., & Kalyanaraman, K. (2011). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *The Review of Economic Studies*, 79(3), 933–959. [26](#), [27](#)
- Lee, M., & Kang, C. (2006). Identification for difference in differences with cross-section and panel data. *Economics Letters*, 92(2), 270–276. [6](#), [8](#)
- Otero, S., & Rau, T. (2017). The effects of drinking and driving laws on car crashes, injuries, and deaths: Evidence from chile. *Accident Analysis Prevention*, 106, 262–274. [20](#), [22](#), [23](#), [24](#), [25](#), [26](#)

A Appendix

A.1 ADDITIONAL FIGURES

Figure A1: DISTRIBUTION OF TEST SCORES



Note: This figure displays a histogram for the distribution of normalized test scores. It shows that they are centered with respect to its mean, so results using this variable as a dependent variable should be interpreted as standard deviations.