Research Question

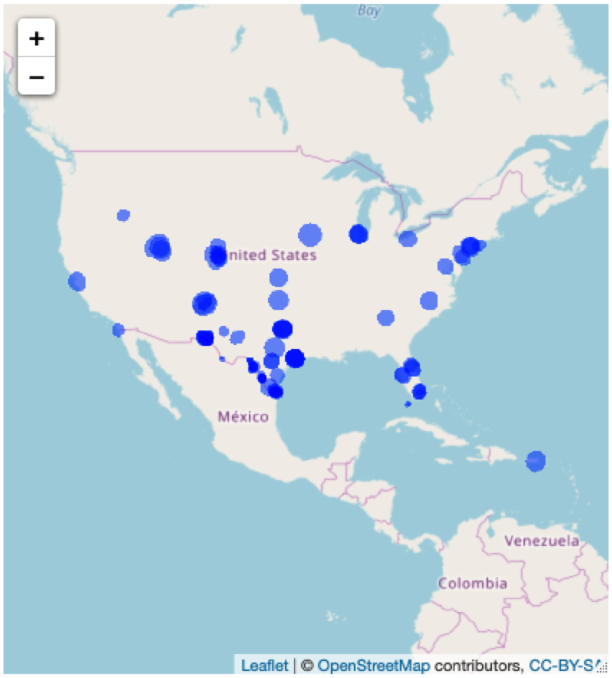
The media we consume is critical to shaping our sense of identity, and prior work has highlighted its importance across domains and in multiple contexts: Oberholzer-Gee, Waldfogel (AER 2009) demonstrate that the presence of Spanish language local news increases Hispanic voter turnout, while Yanigazawa-Drott (QJE 2014) shows that radio broadcasts in Rwanda contributed to the violence and genocide that took place in the 90s.

In the next few pages, I aim to examine the causal effect of Spanish language television (SLTV) on schooling outcomes for Hispanic people. Specifically, I look at the potentially adverse discipline consequences that may arise from the presence of television, ranging from out-of-school suspensions to race/ethnicity-based harassment.

Method and Model

To isolate the causal effect of Spanish language television, I adopt the technique used in Newman, Velez (AJPS 2019) and generalize it from three counties to the entirety of the US. Newman and Velez exploit a FCC (Federal Communications Commission) regulation which determines the distance from a TV station in which the station's broadcast signal is protected from interference. This creates a natural regression discontinuity, where the decaying strength of a signal over distance is combined with this cutoff in broadcast protection to create a split among people just inside and outside these coverage 'contours' that are presumably comparable save for their access to broadcast TV.

Figure 1: The Coverage Contours of Spanish Language TV stations



In the case of Spanish language TV in particular, this should allow me to examine its causal effect on Hispanic populations for spatially located outcomes, such as public schooling results. It's worth noting that these contours are purely determined by an algorithm that looks at things like local elevation and antennae strength, so that the cutoffs are located in more or less random locations, and that coverage is large enough that these contours tend to cut across towns and suburbs, rather than cities.

A standard regression thus looks like restricting the universe of schools to only those within a small radius of the contour boundary, where the key independent variable of interest is an indicator for the school being inside or outside the boundary, interacted with the distance to the boundary:

$$Y_i^{j,k} = \beta_0 + \beta\mathbb{I}[InsideContour_i] \times Distance_i + \gamma X_i + \delta Z^j + \epsilon_i^k \qquad \epsilon \overset{iid}{\sim} N(0, \sigma_i^{k^2})$$

where Y_i is an outcome for school i in county j and school district k, X is a vector of school-level controls, and Z is a vector of county-level controls. Errors are often clustered by school district, meaning that $Corr(\sigma_i^k, \sigma_{i'}^k) \neq 0$ is permissible.

When the outcome variable is a binary variable, the model instead follows:

$$\mathbb{P}(Y_i^j = 1|X, Z) = \frac{\exp[\beta_0 + \beta\mathbb{I}[InsideContour_i] \times Distance_i + \gamma X_i + \delta Z^j]}{1 + \exp[\beta_0 + \beta\mathbb{I}[InsideContour_i] \times Distance_i + \gamma X_i + \delta Z^j]}$$

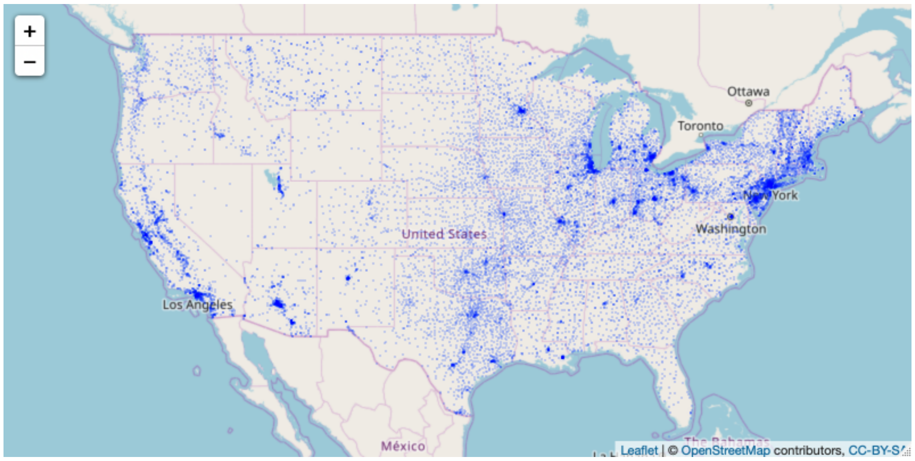
for a logistic/logit regression.

Data

Data for the instrument comes from both the FCC and TMS (a telecommunications company that was kind enough to let me use their API for free). The relevant data here is essentially just the coverage contour spatial data and the broadcast language of the station.

The data on public schools comes from the US government's CRDC (Civil Rights Data Collection) dataset. It's a very large dataset with over 500 outcome/control variables (the vast majority of these are not suitable as controls for one another in this setting), and importantly, it breaks down all major variables of interest by ethnicity. These are all at the school level, and the geographic location of these schools is mapped using ArcGIS.

Figure 2: Map of School Districts in the US



Additional controls like population, income, density of Hispanic population etc. at the county level are from IPUMS.

Some summary statistics of interest are presented below:

Table 1: School-District Level Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Distance to Boundary	17,280	136.855	146.751	0.000	15.786	217.567	806.543
SLTV Coverage Dummy	17,280	0.292	0.455	0.000	0.000	1.000	1.000
% County Hispanic	17,280	7.051	11.950	0.000	0.668	6.974	97.216
Log(Population)	17,280	11.618	1.840	5.869	10.242	13.110	15.997
Log(Income)	17,280	9.428	0.257	7.976	9.257	9.593	10.245

Note: Distance to SLTV Boundary measured in KM

Table 2: School Level Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Total Students	96,349	524.859	449.354	2.000	254.000	662.000	14,164.000
# Hispanic Students	91,019	143.195	243.873	2.000	13.000	166.000	7,675.000
Contains Grade 1	96,350	0.538	0.499	0	0	1	1
Contains Grade 6	96,350	0.364	0.481	0	0	1	1
Contains Grade 9	96,350	0.253	0.435	0	0	1	1
Hispanic Suspension Dummy	94,535	0.382	0.486	0.000	0.000	1.000	1.000
Hispanic Harassment Victim Dummy	94,127	0.026	0.160	0.000	0.000	0.000	1.000
Hispanic Harassment Offender Dummy	94,354	0.023	0.149	0.000	0.000	0.000	1.000
# Teachers	93,934	35.219	33.892	1.000	19.000	44.000	6,031.000

Note: Dummies indicate whether event occurred in the school over the past year

Suspensions

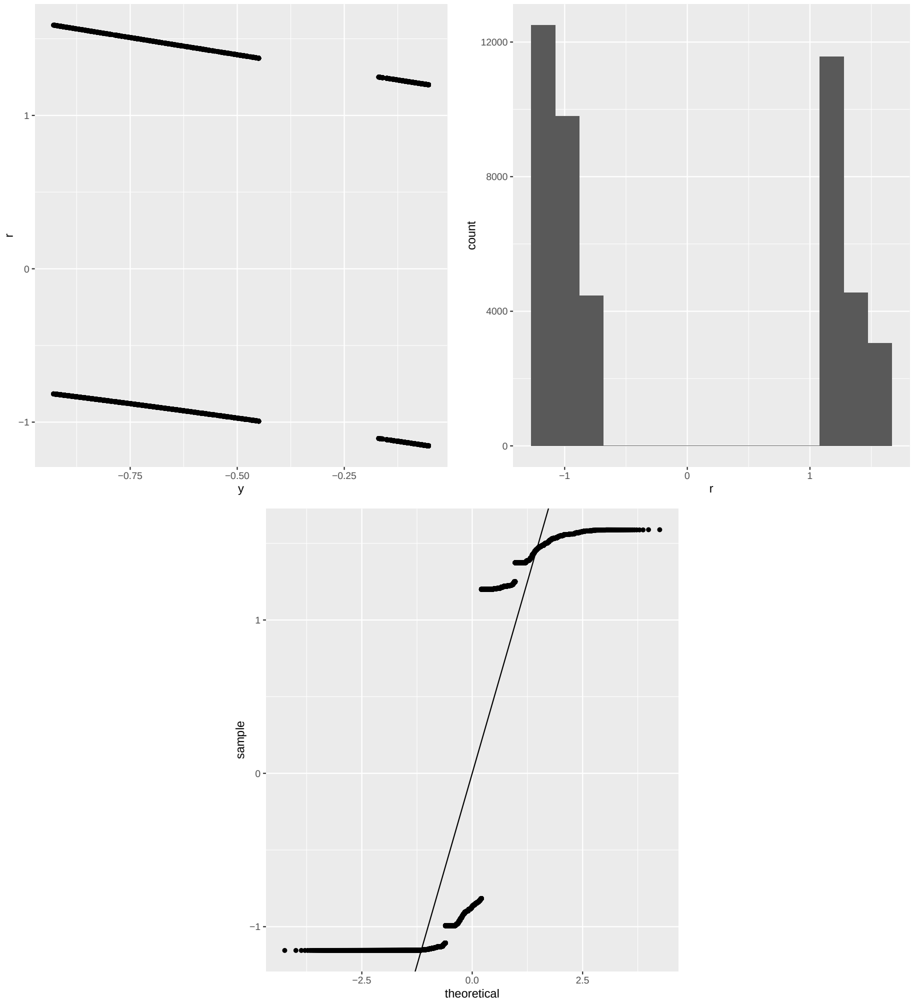
In the following regressions, I examine out of school suspensions. First, with a logit approach using a dummy for a school in the past year ever having a Hispanic student receiving at least one out of school suspension:

Table 3: Effect of TV on Hispanic Out of School Suspension Dummy

	Dependent variable:				
	Dummy for Hispanic Out of School Suspension				
	(1)	(2)	(3)	(4)	(5)
TV Dummy	0.397*** (0.027)	0.092*** (0.030)	0.204*** (0.031)	0.064* (0.033)	−0.006 (0.035)
TV Dummy × Distance to Boundary	0.003*** (0.001)	0.006*** (0.001)	0.005*** (0.001)	0.004*** (0.001)	0.005*** (0.001)
Distance to Boundary (meters)	−0.005*** (0.0004)	−0.004*** (0.0004)	−0.004*** (0.0004)	−0.004*** (0.0005)	−0.003*** (0.0005)
Log(Population)		0.074*** (0.007)	0.138*** (0.008)	0.135*** (0.009)	0.102*** (0.010)
% County Hispanic		1.714*** (0.069)	1.127*** (0.081)	1.210*** (0.088)	−1.383*** (0.109)
Log(Income)			−0.664*** (0.046)	−1.180*** (0.050)	−1.024*** (0.054)
# Teachers at School				0.031*** (0.0005)	0.010*** (0.001)
# Hispanic Students					0.005*** (0.0001)
Total Students					0.0004*** (0.0001)
Contains Grade 1					−0.887*** (0.027)
Contains Grade 6					0.299*** (0.024)
Contains Grade 9					0.126*** (0.031)
Observations	45,947	45,947	45,947	45,947	45,947
Log Likelihood	−30,733.950	−30,315.250	−30,211.380	−27,500.700	−24,898.820
Akaike Inf. Crit.	61,475.890	60,642.500	60,436.760	55,017.410	49,823.650
Note:	*p<0.1; **p<0.05; ***p<0.01				

Column 1 presents the most simple model without any additional controls. The figure below presents some regression diagnostics (the fitted outcome against studentized residuals, a histogram of the studentized residuals, and a QQ plot of the studentized residuals):

Figure 3: Table 3, Column 1 Regression Diagnostics

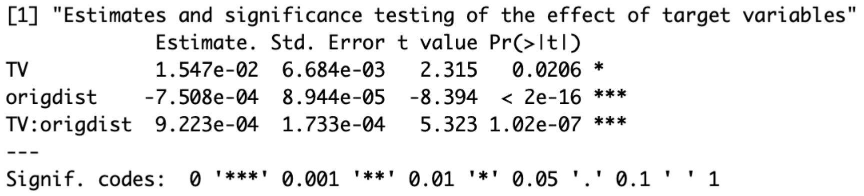


It is clear that are outstanding issues with the regression: they are clumped into two groups based on whether the outcome holds or not. This indicates that there are still large underlying sources of variation that are not being captured by the regression itself. Given the spatial nature of the identification method, it is natural to control for the demographic features of the areas in which the schools are located, which is what is done in columns (2) and (3).

The key result, that television drives negative outcomes, is nonetheless present in this first regression: this is visible from the positive television dummy, but also from the interaction term. If there is indeed an effect from television, we would expect these to be significant in the same direction, as the dummy captures the FCC regulation imposed cutoff, while there is also a natural decay in signal strength over distance.

Column (5) adds in school-level controls. However, the universe of these is large, and they are potentially collinear. Thus, the variables present in the regression are selected using forward stepwise BIC. While the partial F test is inappropriate for this context (logit), employing a likelihood ratio test also justifies each successive addition of variables across the columns of this regression.

Figure 4: Table 3, Column 5 Regression Diagnostics

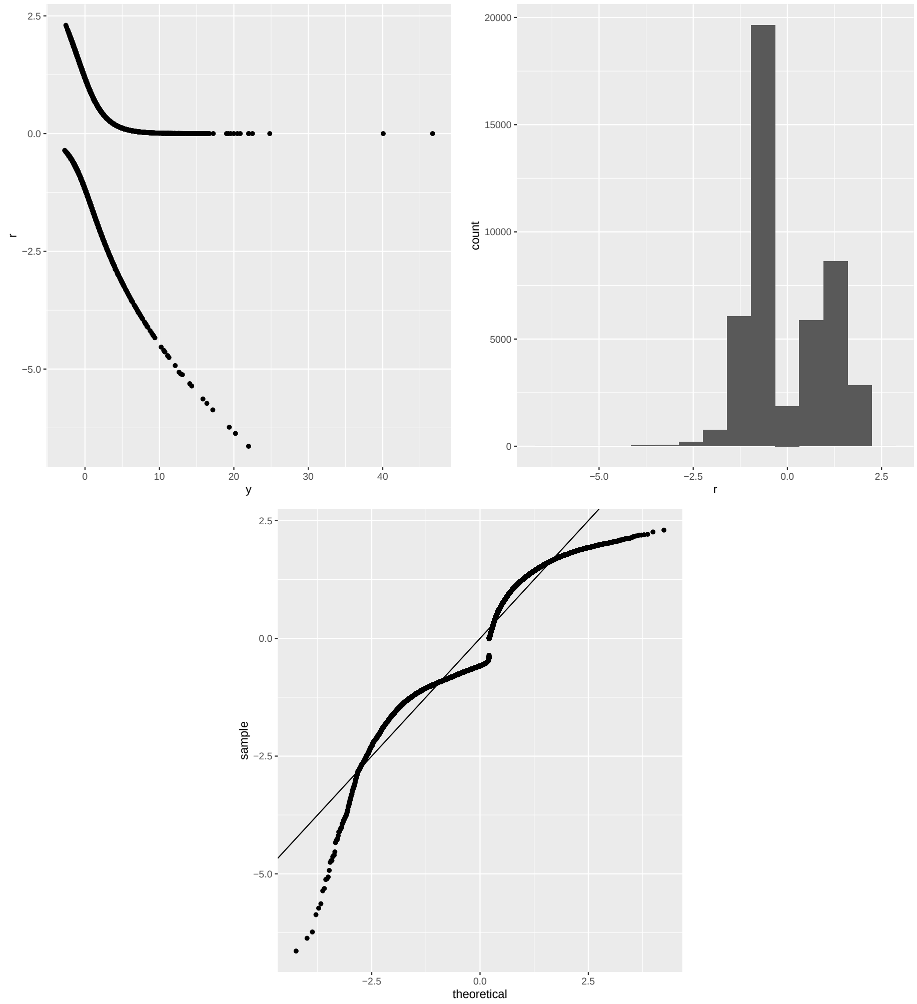


```
[1] "Estimates and significance testing of the effect of target variables"
             Estimate. Std. Error t value Pr(>|t|)
TV           1.547e-02  6.684e-03   2.315   0.0206 *
origdist    -7.508e-04  8.944e-05  -8.394  < 2e-16 ***
TV:origdist  9.223e-04  1.733e-04   5.323 1.02e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, a regression utilizing LASSO is presented; the information matches what we saw in the prior regressions, with the effect of television and the interaction both being positive and significant for $\alpha = .1$.

To verify that these additional controls have mitigated the concerns of non-normal residuals noted above, the diagnostics for the regression in column (5) are presented:

Figure 5: Table 3, Column 5 Regression Diagnostics



There are still two apparent clusters, but they are substantially less differentiated than in the simplest regression above. It's worth noting that most of the progress is made with the addition of county-level variables. Nonetheless, it still appears as if there is non-constant variance and an overall non-normal distribution. These problems can likely be attributed to the large number of 0s present in the regression (close to 2/3 of values). Thus, a zero-inflated model with logit as the link and a Poisson distribution is used too, with results verified:

Table 4: Effect of TV on Hispanic Out of School Suspension Dummy, Zero-Inflated

	Dependent variable:		
	# Hispanic Out of School Suspensions		
	(1)	(2)	(3)
TV Dummy	0.443^{***}	0.151^{***}	0.189^{***}
	(0.007)	(0.008)	(0.008)
TV Dummy × Distance to Boundary	0.003^{***}	0.006^{***}	0.003^{***}
	(0.0002)	(0.0002)	(0.0002)
Distance to Boundary (meters)	-0.004^{***}	-0.005^{***}	-0.004^{***}
	(0.0001)	(0.0002)	(0.0002)
Log(Population)		0.160^{***}	0.131^{***}
		(0.002)	(0.002)
% County Hispanic		0.735^{***}	-0.215^{***}
		(0.018)	(0.018)
Log(Income)		-0.508^{***}	-0.654^{***}
		(0.012)	(0.013)
# Teachers at School			0.004^{***}
			(0.00001)
# Hispanic Students			0.001
Total Students			−0.0003
Contains Grade 1			-1.072^{***}
			(0.002)
Contains Grade 6			0.053^{***}
			(0.005)
Contains Grade 9			0.034^{***}
			(0.005)
Observations	45,947	45,947	45,947
Log Likelihood	−178,855.800	−169,849.100	−116,549.100
Note:			*p<0.1; **p<0.05; ***p<0.01

From this, we glean that it is plausible for there to be a link between the presence of television and students acting out in ways leading to suspension.