Regression Discontinuity

Data Science for Spatial Systems

Ollie Ballinger

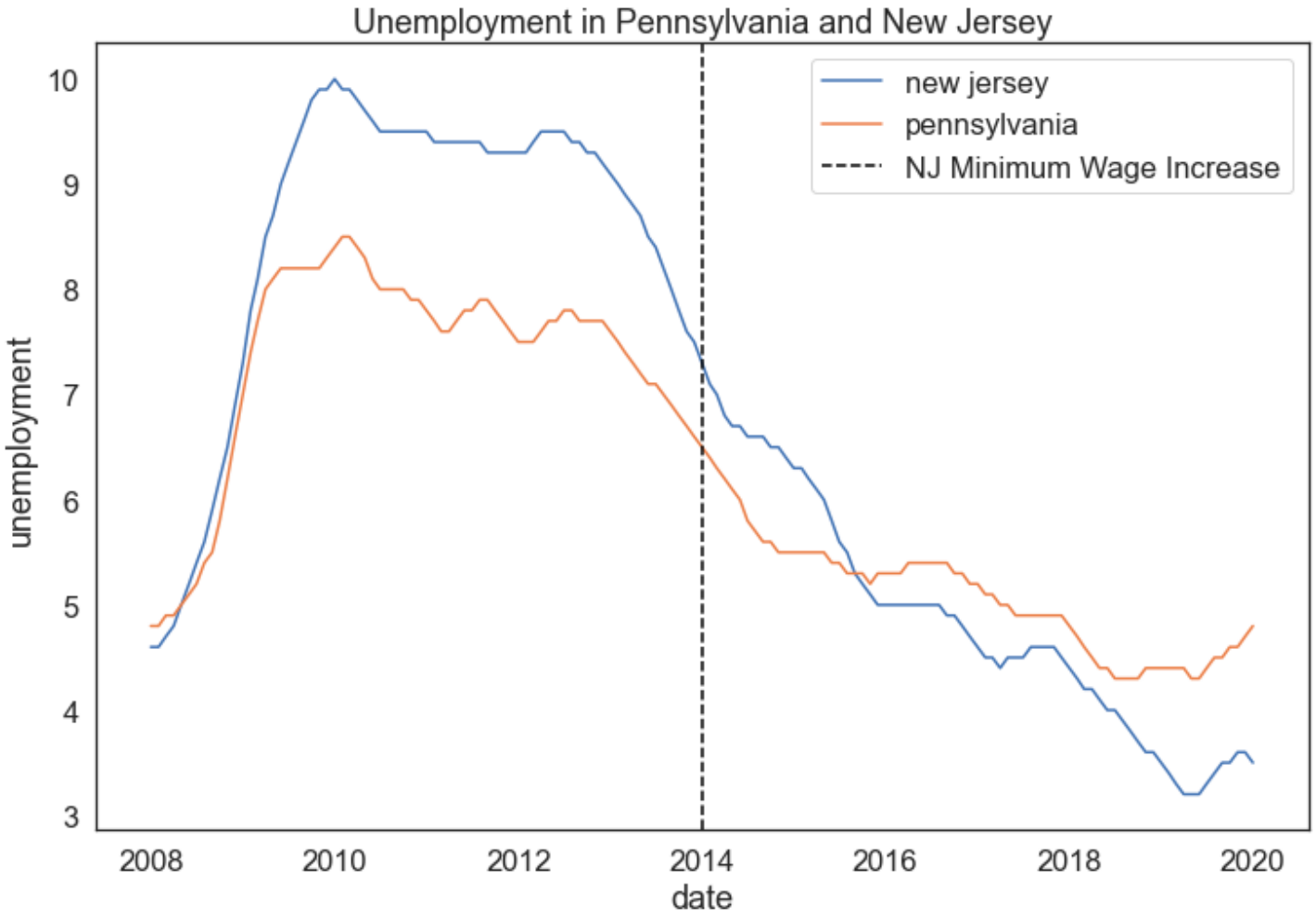
Outline

1. Regression Discontinuity
2. Worked Example

Recap: Difference in Differences

$$Y_{it} = \beta_0 + \beta_1 Post_t + \beta_2 Treat_i + \beta_3 Treat_i \times Post_t + \varepsilon_{it}$$

	Control	Treatment
Before	β_0	$\beta_0 + \beta_2$
After	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$



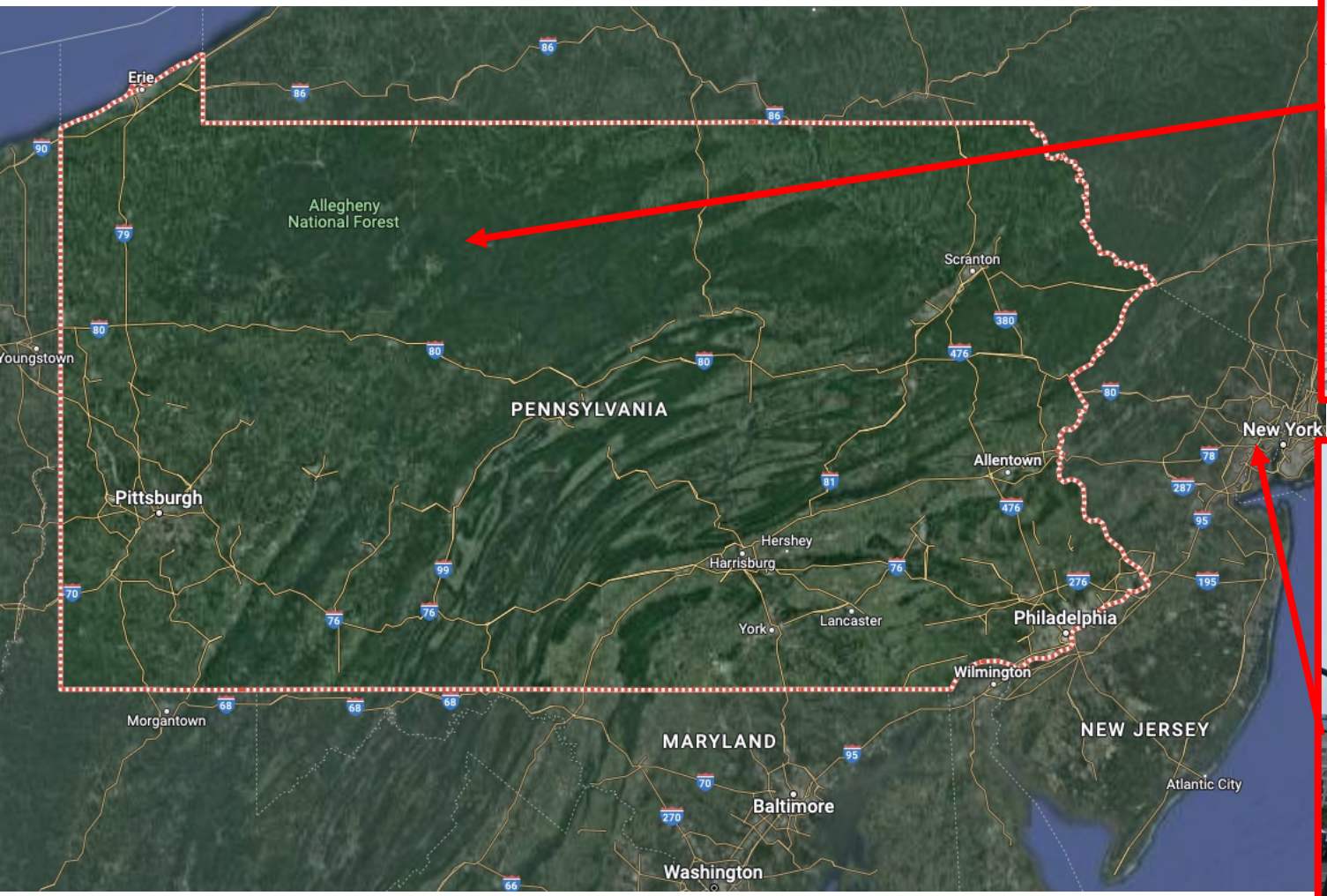
Regression Discontinuity

Counterfactuals

- Difference in differences requires a few things:
  - A treatment located at a point in time
  - Two distinct groups for which we have measurement pre-and post-treatment
  - Parallel pre-treatment trends in the outcome variable y
  - No simultaneous treatment occurring around our treatment of interest
- These allow us to construct a valid **counterfactual:**
  - We can argue that the control group in the post-treatment period acts as a valid representation of the treatment group's behaviour in the absence of treatment. We can thus interpret the difference between the two as the causal effect of the treatment.
- But this isn't the only way of constructing a valid counterfactual

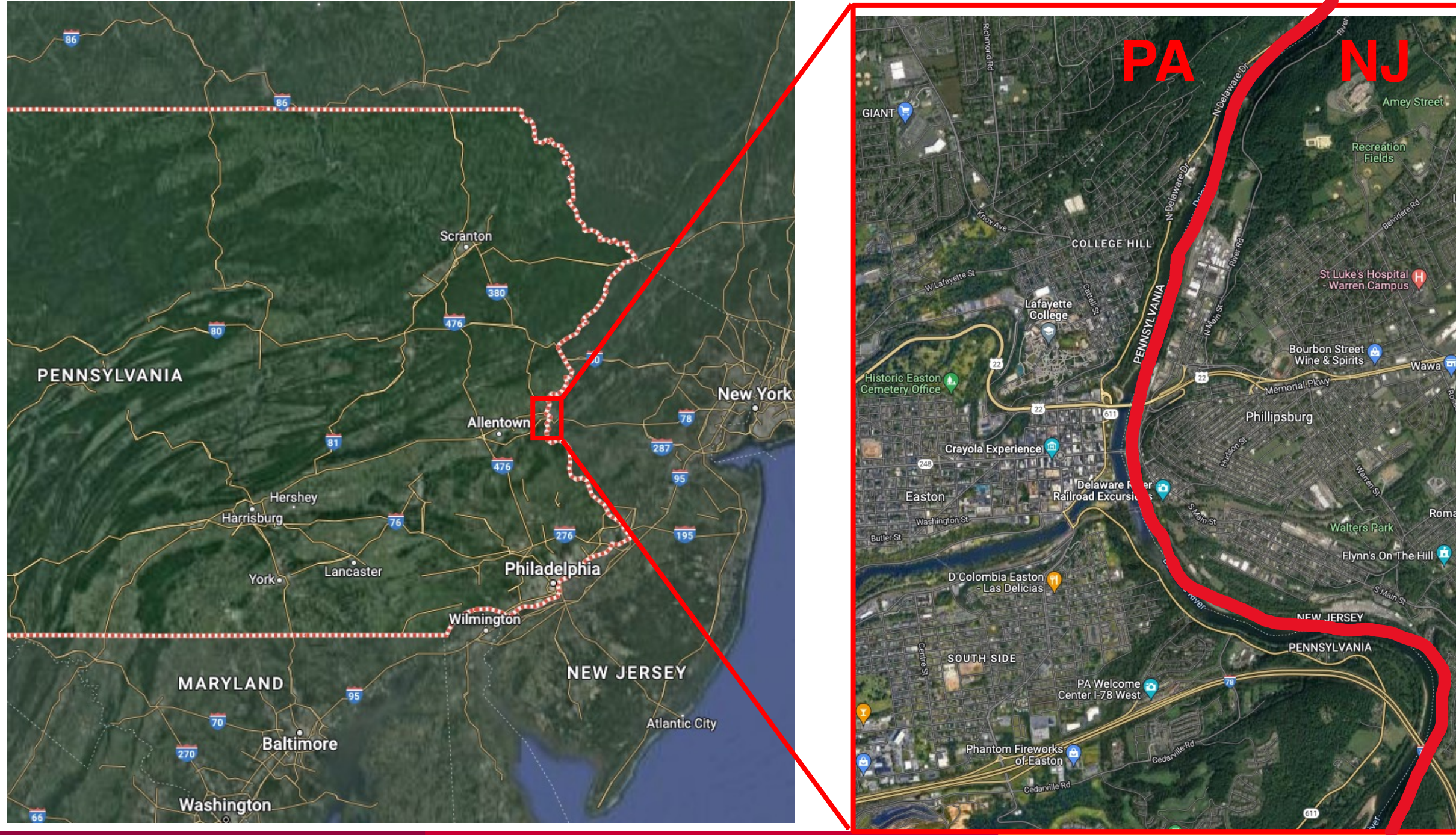
Counterfactuals

- **What if we didn't have pre-treatment data for NJ and PA, but we had county level data on poverty for the period after the minimum wage was introduced.**
- We can't do DiD, since we can't show that NJ and PA had similar trends in poverty pre-treatment. Maybe poverty in NJ was already falling rapidly before they introduced the minimum wage, while it was increasing in PA.
- We also can't just compare all the counties in PA against the counties in NJ, since there are huge states many differences between them that have nothing to do with the minimum wage law.


Erie
Allegheny
National Forest
Scranton
Youngstown
PENNSYLVANIA
New York
Allentown
Pittsburgh
Hershey
Harrisburg
York
Lancaster
Philadelphia
Wilmington
Morgantown
MARYLAND
NEW JERSEY
Atlantic City
Baltimore
Washington


Tiona, PA


Jersey City, NJ


PA
NJ
PENNSYLVANIA
New York
Allentown
Scranton
Hershey
Harrisburg
Lancaster
York
Philadelphia
Wilmington
MARYLAND
NEW JERSEY
Atlantic City
Baltimore
Washington
COLLEGE HILL
Lafayette College
Easton
Phillipsburg
Crayola Experience
Historic Easton Cemetery Office
Bourbon Street Wine & Spirits
Wawa
St Luke's Hospital - Warren Campus
Amey Street
Recreation Fields
Walters Park
Flynn's On The Hill
D Colombia Easton - Las Delicias
PA Welcome Center I-78 West
Phantom Fireworks of Easton
SOUTH SIDE
NEW JERSEY
PENNSYLVANIA

Regression Discontinuity Design

- Regression Discontinuity Design (RDD) is a quasi-experimental evaluation option that measures the impact of an intervention, or treatment, by applying a treatment assignment mechanism based on a cutoff point in a continuous eligibility index.
- RDD estimates local average treatment effects around the cutoff point, where treatment and comparison units are most similar.

Source: World Bank

RDD Framework

- **Question**: Do minimum wages increase unemployment?
- **Running variable**:
  - State– NJ and PA are hugely different beyond the fact that one implemented a minimum wage policy and the other didn't
- **Exogenous cutoff**:
  - State border
- **Bandwidth**:
  - If we compare areas in PA and NJ within a small distance of the border between these states, they are probably going to be very similar in terms of demographics, economic structure, etc.

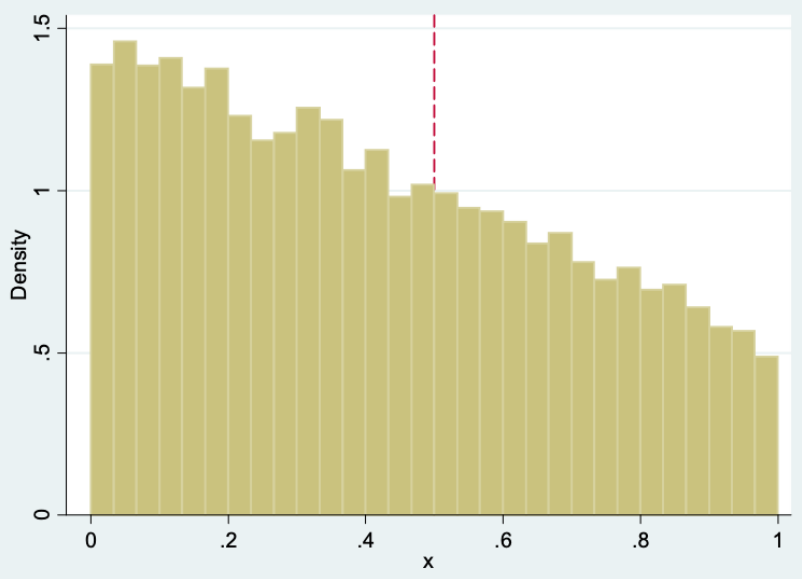
Conditions

1. A continuous eligibility index
   - A continuous measure on which the population of interest is ranked (i.e. test score, poverty score, age).
2. A clearly defined cutoff point
   - A point on the index above or below which the population is determined to be eligible for the program.
   - For example, students with a test score of at least 80 of 100 might be eligible for a scholarship, households with a poverty score less than 60 out of 100 might be eligible for food stamps, and individuals age 67 and older might be eligible for pension. The cutoff points in these examples are 80, 60, and 67, respectively. The cutoff point may also be referred to as the threshold.

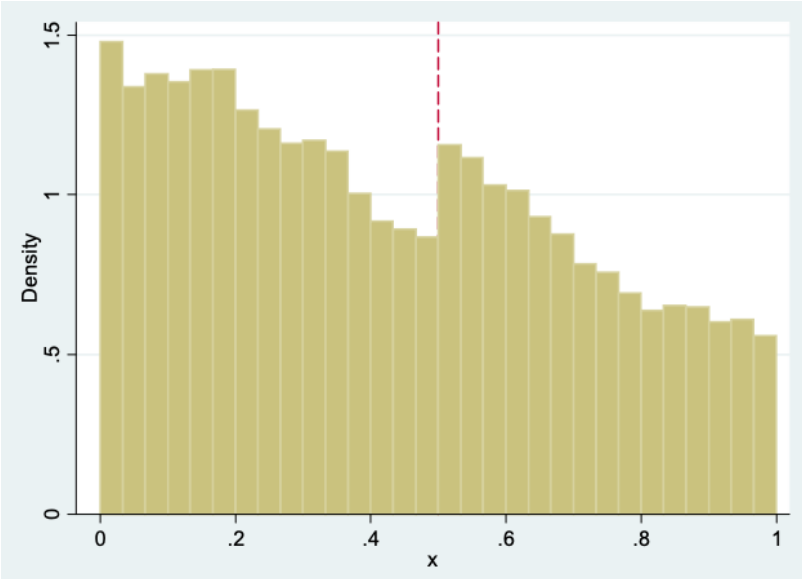
Assumption 1

The eligibility index should be continuous around the cutoff point. There should be no jumps in the eligibility index at the cutoff point or any other sign of individuals manipulating their eligibility index in order to increase their chances of being included in or excluded from the program.

Continuous distribution



Heaping around the cutoff



Assumption 2

- Individuals close to the cutoff point should be very similar, on average, in observed and unobserved characteristics.
  - In the RDD framework, this means that the distribution of the **observed and unobserved variables should be continuous around the threshold.**
  - Even though researchers can check similarity between observed covariates, the similarity between unobserved characteristics must be assumed. This is considered a plausible assumption to make for individuals very close to the cutoff point, that is, for a relatively narrow window.

Original RDD paper

- Thistlethwaite and Campbell (1960) study the effects of college scholarships on later students' achievements

- Scholarships are granted based on whether a student's test score exceeds some threshold *c*

- Consider the following variables:
  - Binary treatment D is receipt of scholarship
  - Covariate *X* is test score with threshold *c*
  - Outcome Y is subsequent earnings
  - Y_0 denotes potential earnings without the scholarship
  - Y_1 denotes potential earnings with the scholarship

Source: Jonathan Mummolo

Original RDD Paper

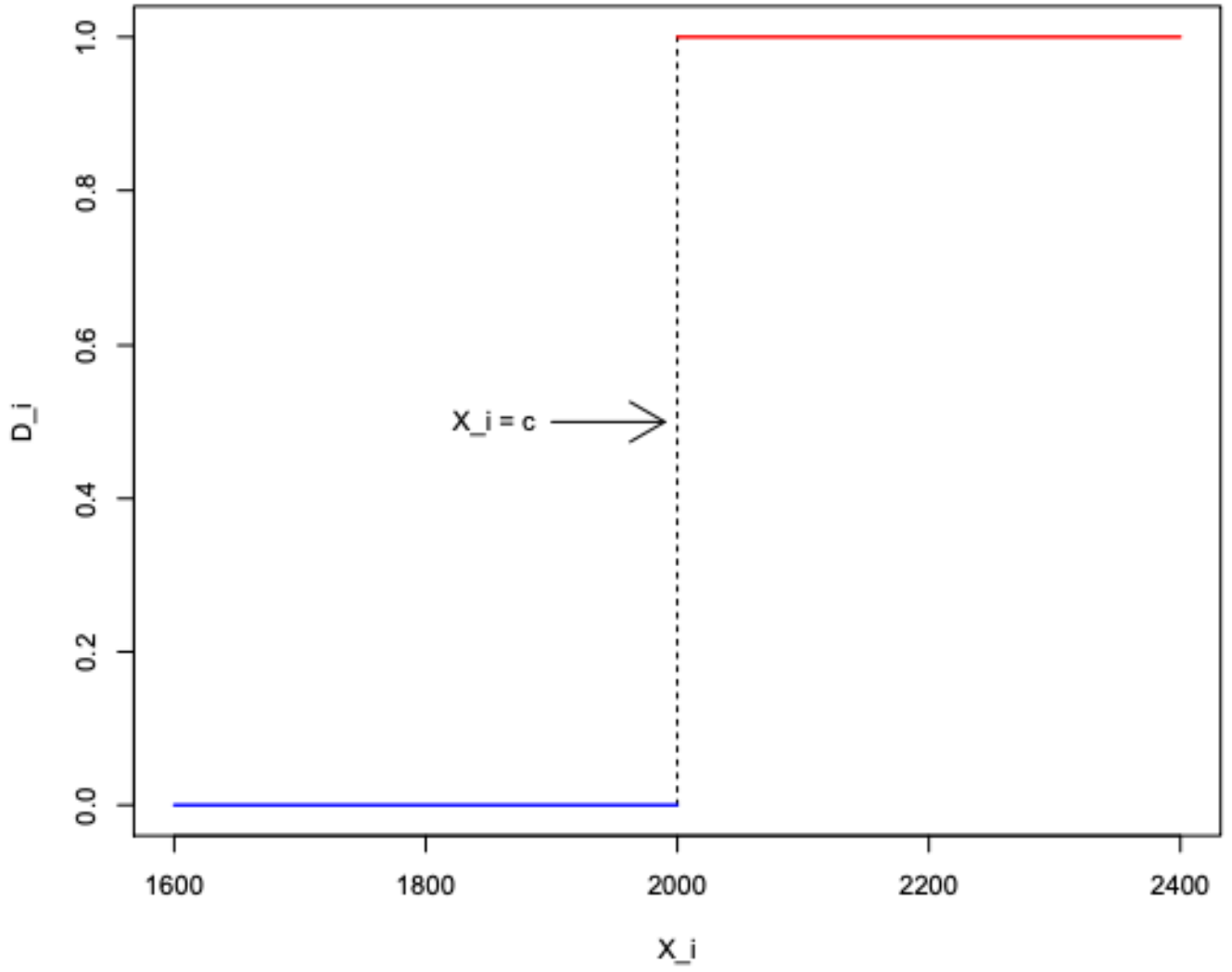
- Assignment to the scholarship treatment Di is completely determined by the value of the SAT score X_i being on either side of the threshold c:

$$D_i = 1\{X_i > X_c\}$$

- X is called the forcing variable, because it "forces" units from control into treatment once X_i exceeds c
- X may be correlated with Y_0 and Y_1 so comparing treated and untreated units does not provide causal estimates (e.g. students with higher SAT scores obtain higher earnings even without the scholarship)
- If the relationship between X and the potential outcomes Y_1 and Y_0 is "smooth" around the threshold c, we can use the discontinuity created by the treatment to estimate the effect of D on Y at the threshold

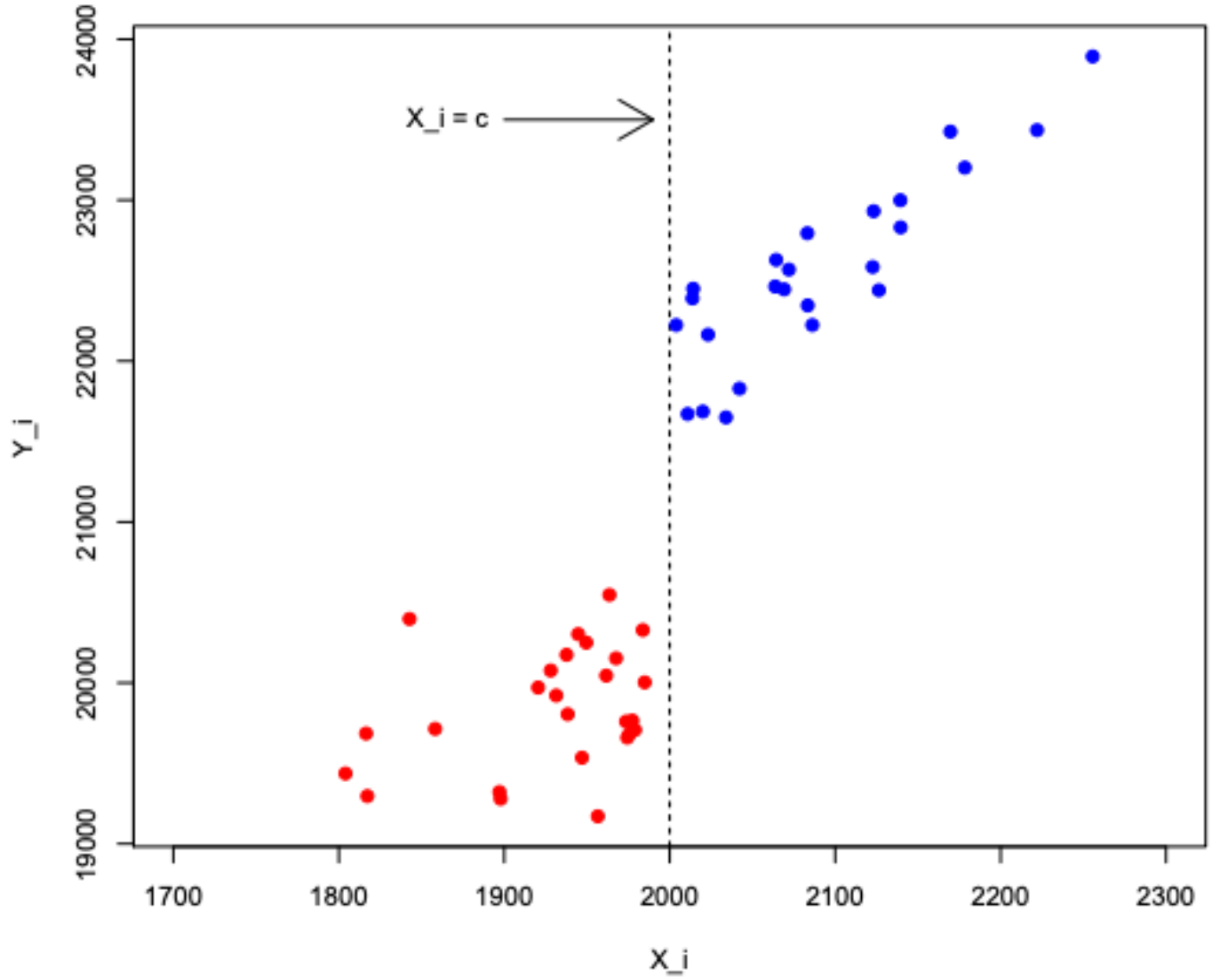
Treatment Assignment

$$D_i = \begin{cases} 1 \; if \; X_i > X_c \\ 0 \; if \; X_i < X_c \end{cases}$$



Observed Outcomes

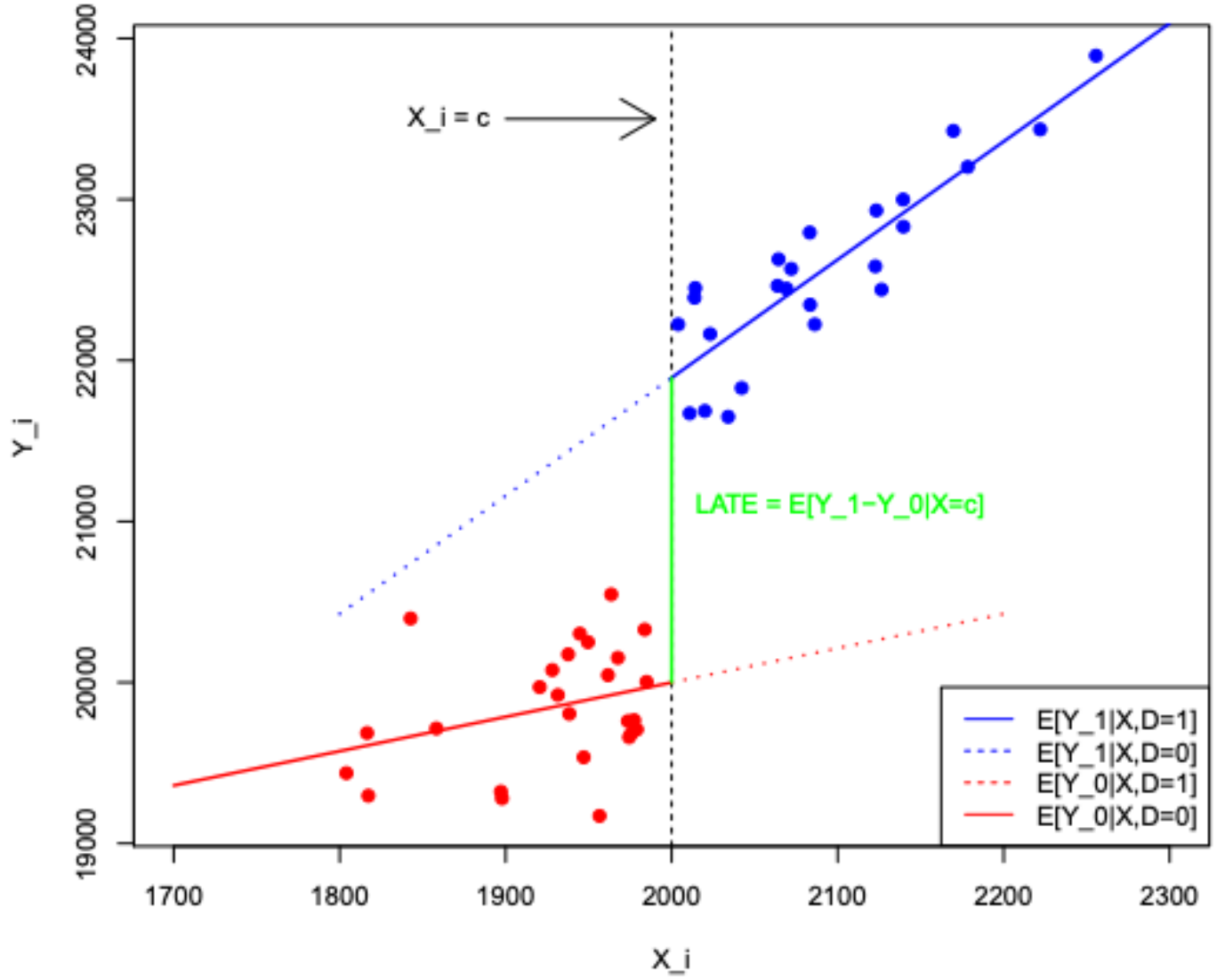
- There appears to be a jump in the observed values of the outcome variable around the cutoff



Potential Outcomes

- We can use the discontinuity created by the treatment to estimate the effect of D on Y at the threshold:

$$\begin{aligned}\alpha SRDD &= E[Y_1 - Y_0 | X = c] \\ &= E[Y_1 | X = c] - E[Y_0 | X = c] \\ &= \lim_{x \downarrow c} E[Y_1 | X = c] - \lim_{x \uparrow c} E[Y_0 | X = c]\end{aligned}$$



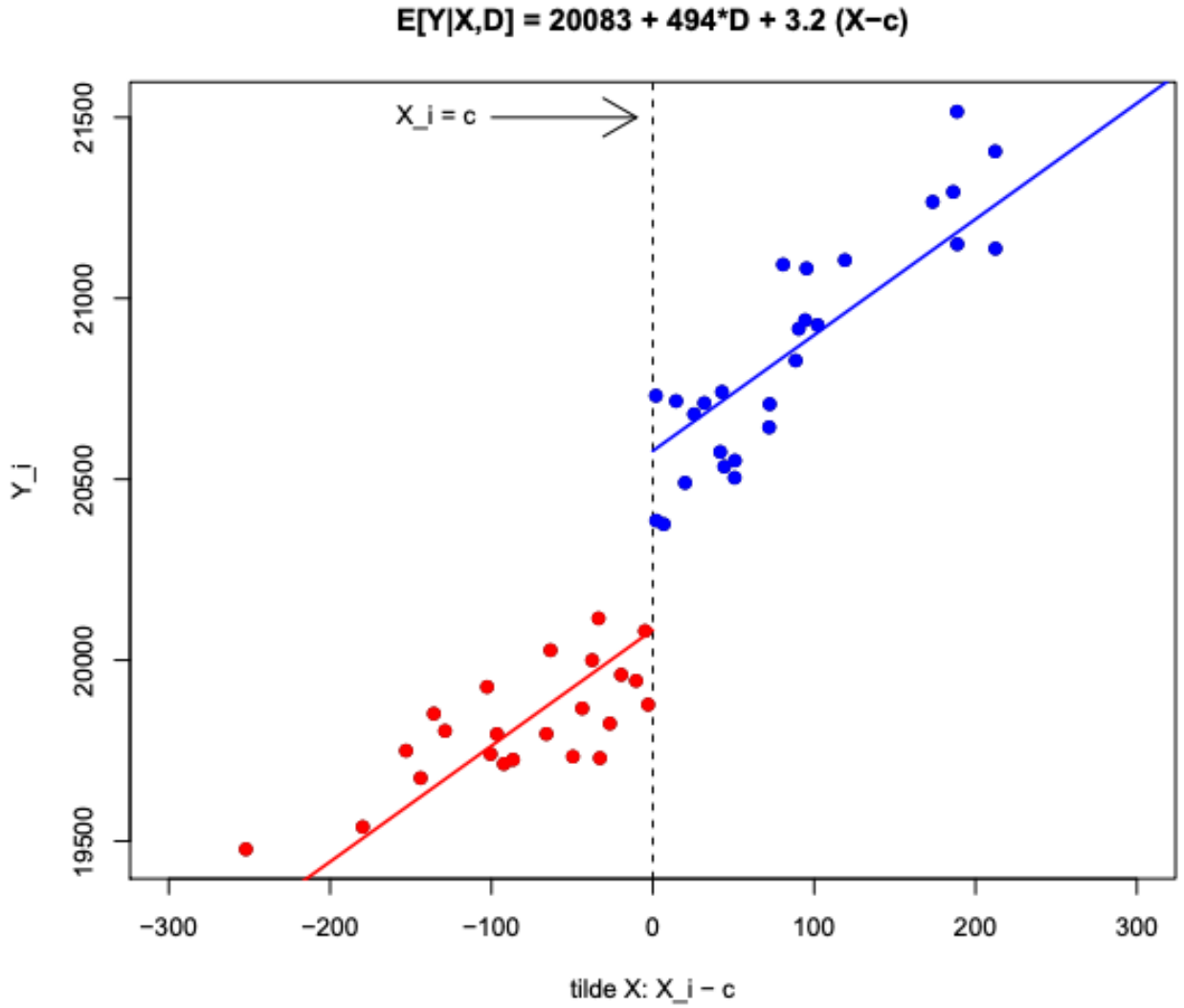
Estimation

1. Trim the sample to a reasonable window around the cutoff c (discontinuity sample):
   - $c - h \leq X_i \leq c + h$, where h is some positive value that determines the size of the window h may be determined by cross-validation
2. Code the margin $\tilde{X}$ which measures the distance to the threshold:
   - $\tilde{X} = X - c$ such that $\tilde{X}_i = \begin{cases} \tilde{X} > 0 \; if \; X > c \;\; and \; thus \; D = 1 \\ \tilde{X} < 0 \; if \; X < c \;\; and \; thus \; D = 0 \end{cases}$
3. Decide on a model for $E[Y \,|X]$
   1. linear, same slope for $E[Y_0|X]$ and $E[Y_1|X]$
   2. linear, different slopes
   3. non-linear

Linear, Same Slopes

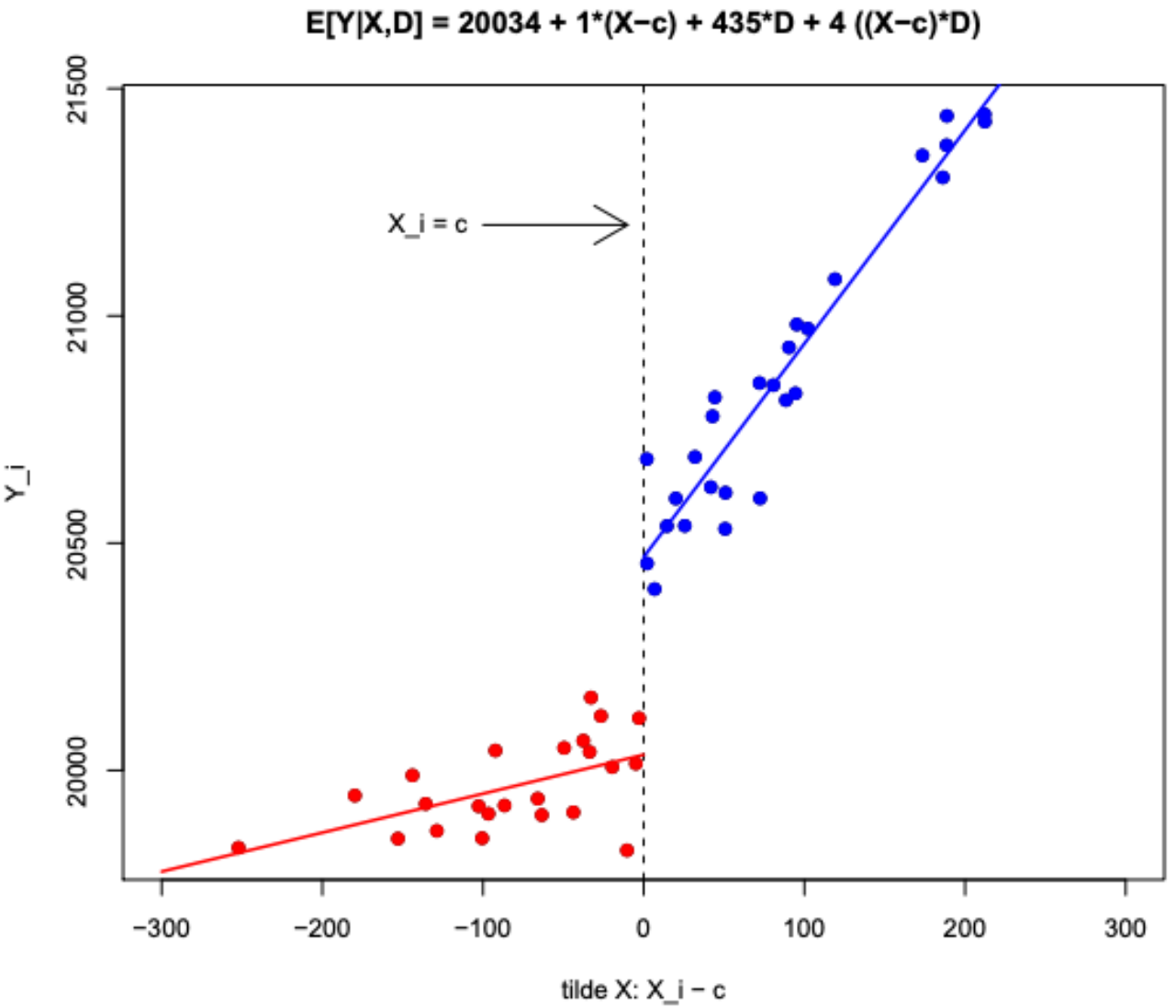
$$Y_i = \gamma + \alpha D_i + \beta \tilde{X}_i + \varepsilon_i$$

$$D_i = \begin{cases} 1 \; if \; \tilde{X} > 0 \\ 0 \; if \; \tilde{X} < 0 \end{cases}$$



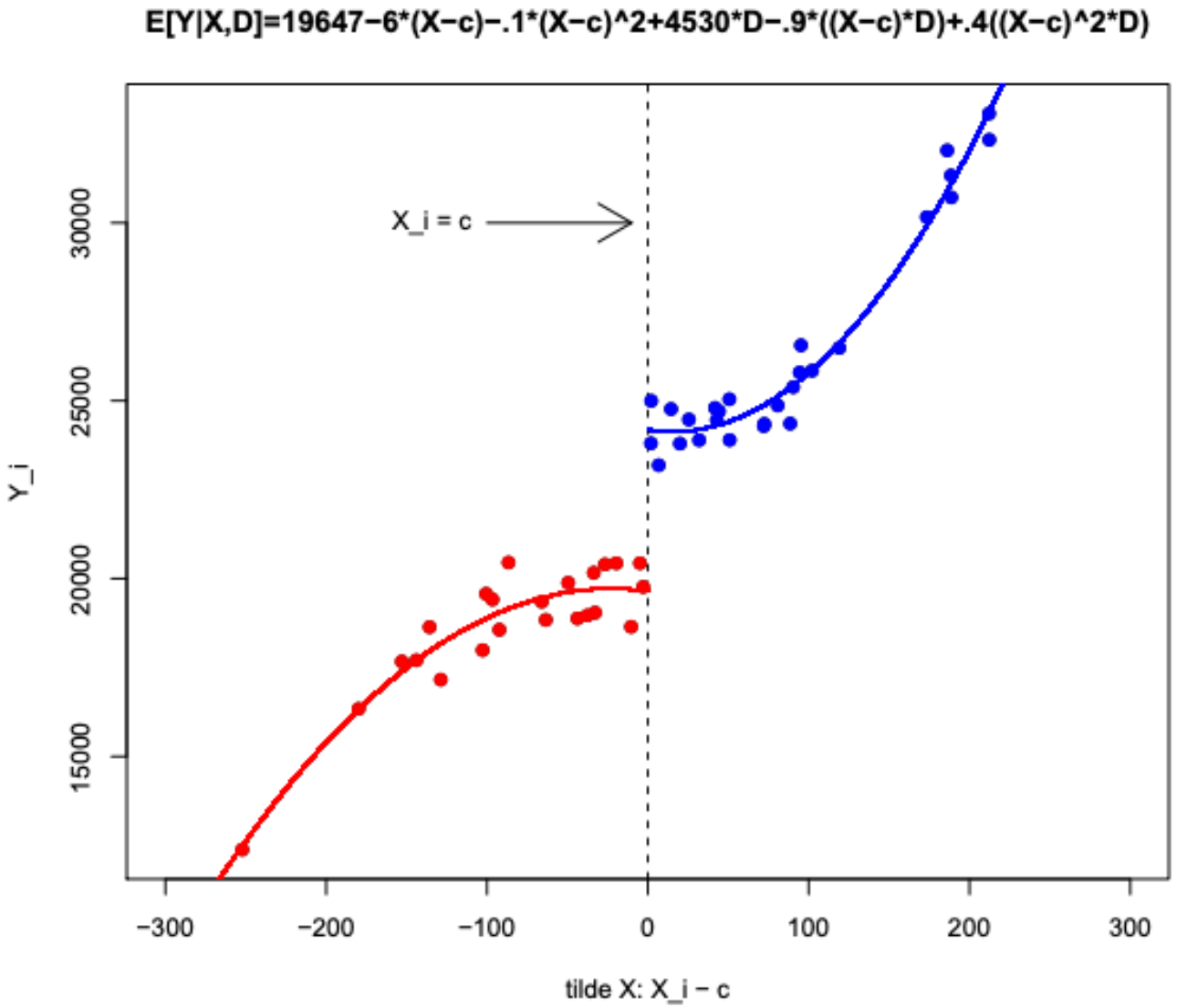
Linear, Different Slopes

$$Y_i = \gamma + \alpha D_i + \beta_0 \tilde{X}_i + \beta_1 (\tilde{X}_i \cdot D_i) + \varepsilon_i$$

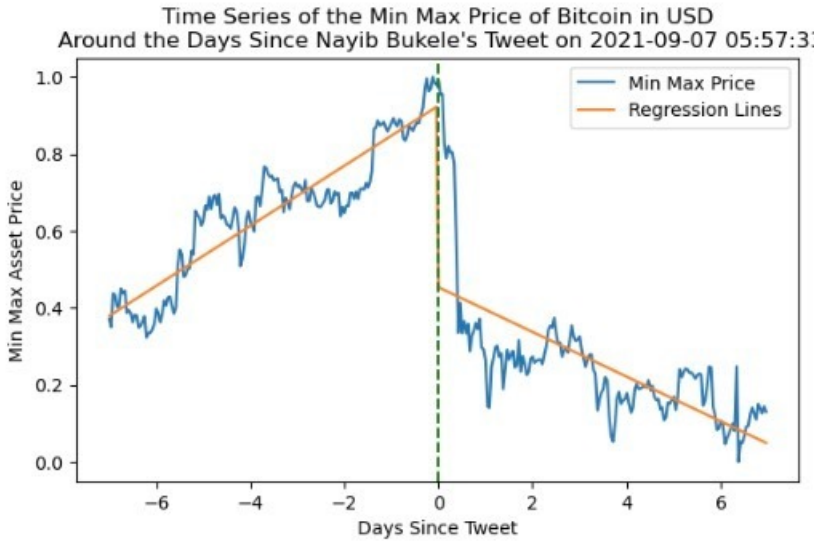


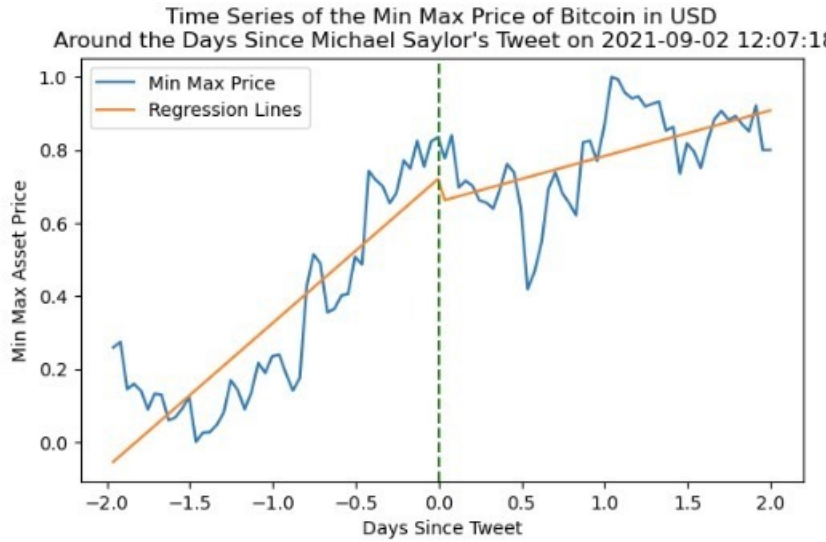
Non-Linear

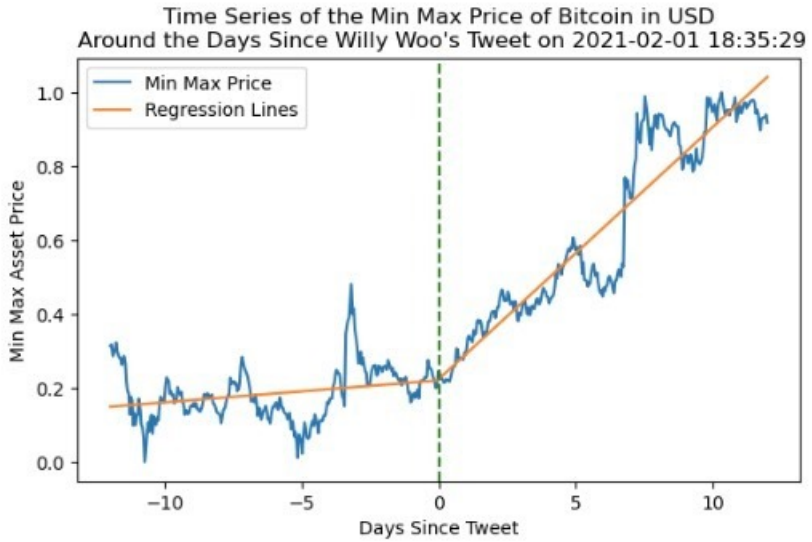
$$Y_i = \gamma_0 + \gamma_1 \tilde{X}_i + \gamma_2 \tilde{X}_i^2 + \alpha_0 D_i + \alpha_1 \left(\tilde{X}_i \cdot D_i\right) + \alpha_2 \left(\tilde{X}_i^2 \cdot D_i\right) + \varepsilon_i$$

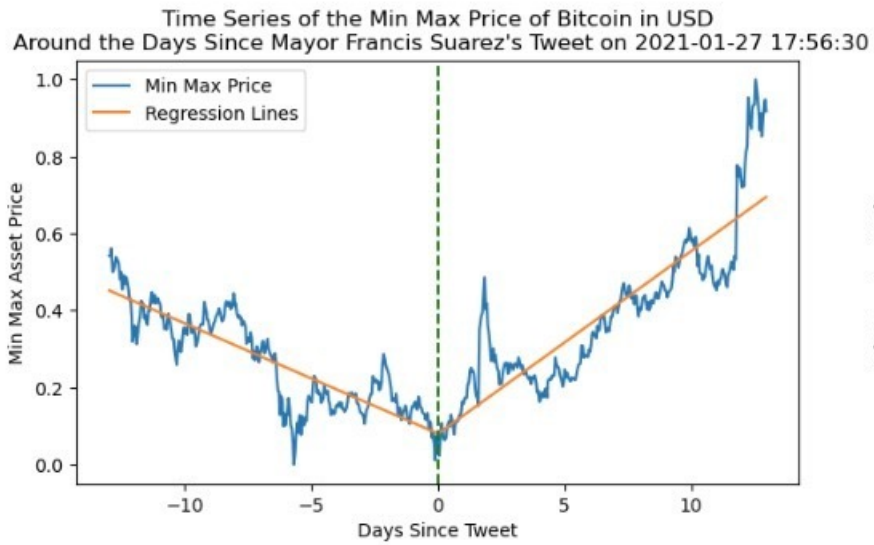


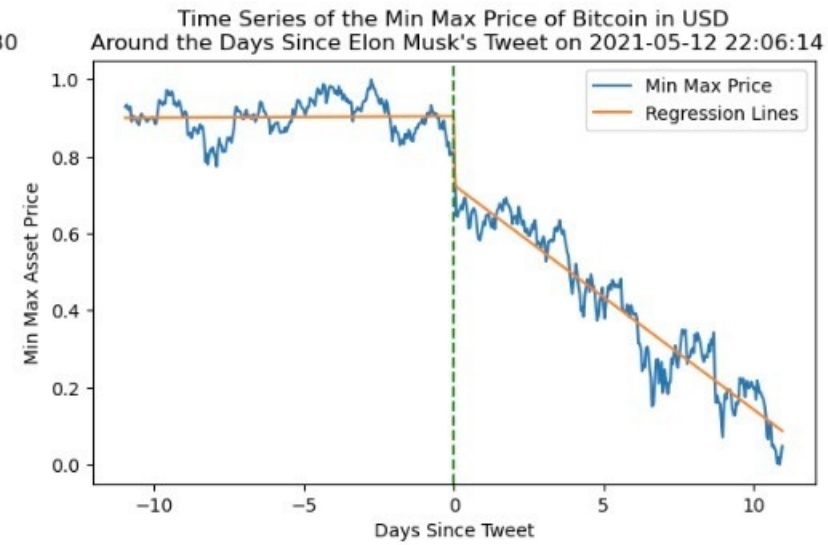
Time Series of the Min Max Prices of each Tweet's Discontinuity Regression Model

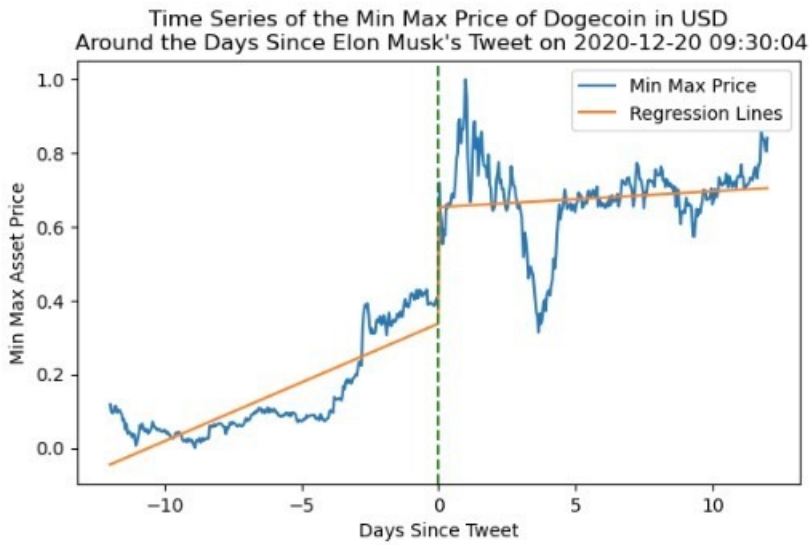








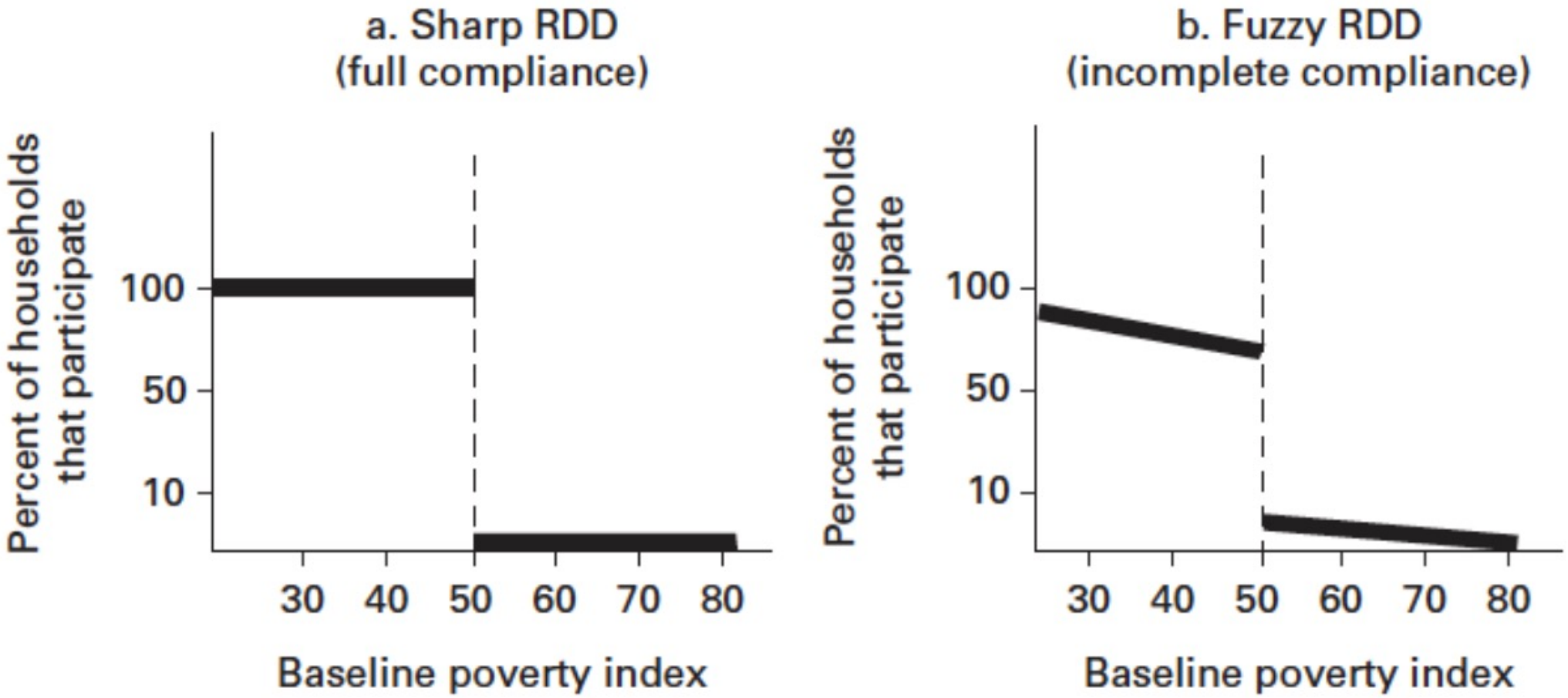




Fuzzy Regression Discontinuity

- Threshold may not perfectly determine treatment exposure, but **it creates a discontinuity in the probability of treatment exposure**
- Incentives to participate in a program may change discontinuously at a threshold, but the incentives are not powerful enough to move all units from non-participation to participation
- We can use such discontinuities to produce instrumental variable estimators of the effect of the treatment (close to the discontinuity)

Sharp vs. Fuzzy RDD



Assigned versus observed treatment

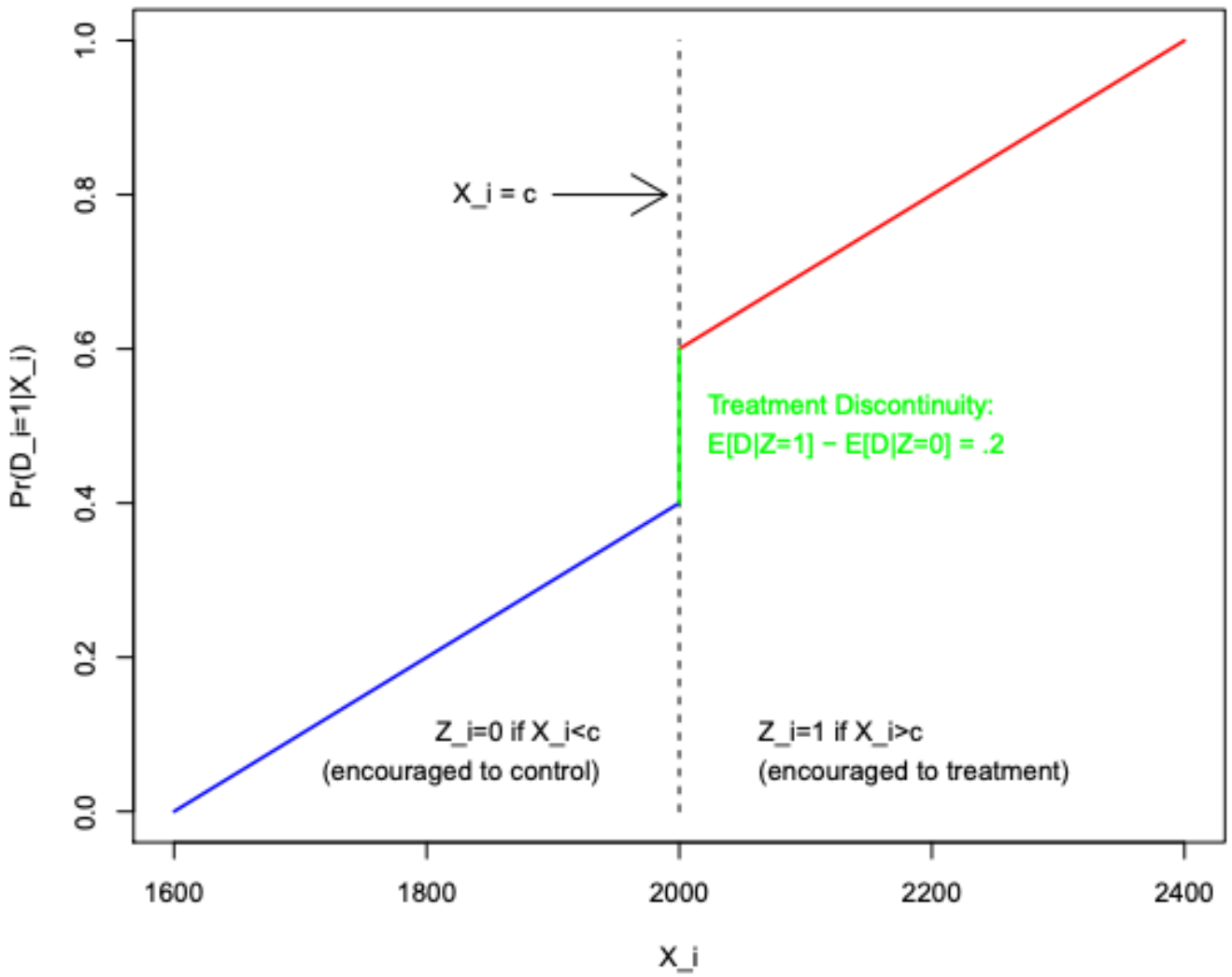
- Assume the treatment is offered to everybody above c, but not everybody might take it
- Let $Z = 1\{X > c\}$ be a binary encouragement indicator that captures whether units are above or below the threshold c
- Let D be the binary observed treatment indicator that captures whether individuals take the treatment or not
- Observed treatment D_i can be modeled as a function of the running variable X_i and the binary encouragement indicator Z_i

$$D_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \varepsilon_i$$

First stage

- Estimating uptake based on assignment to treatment group:

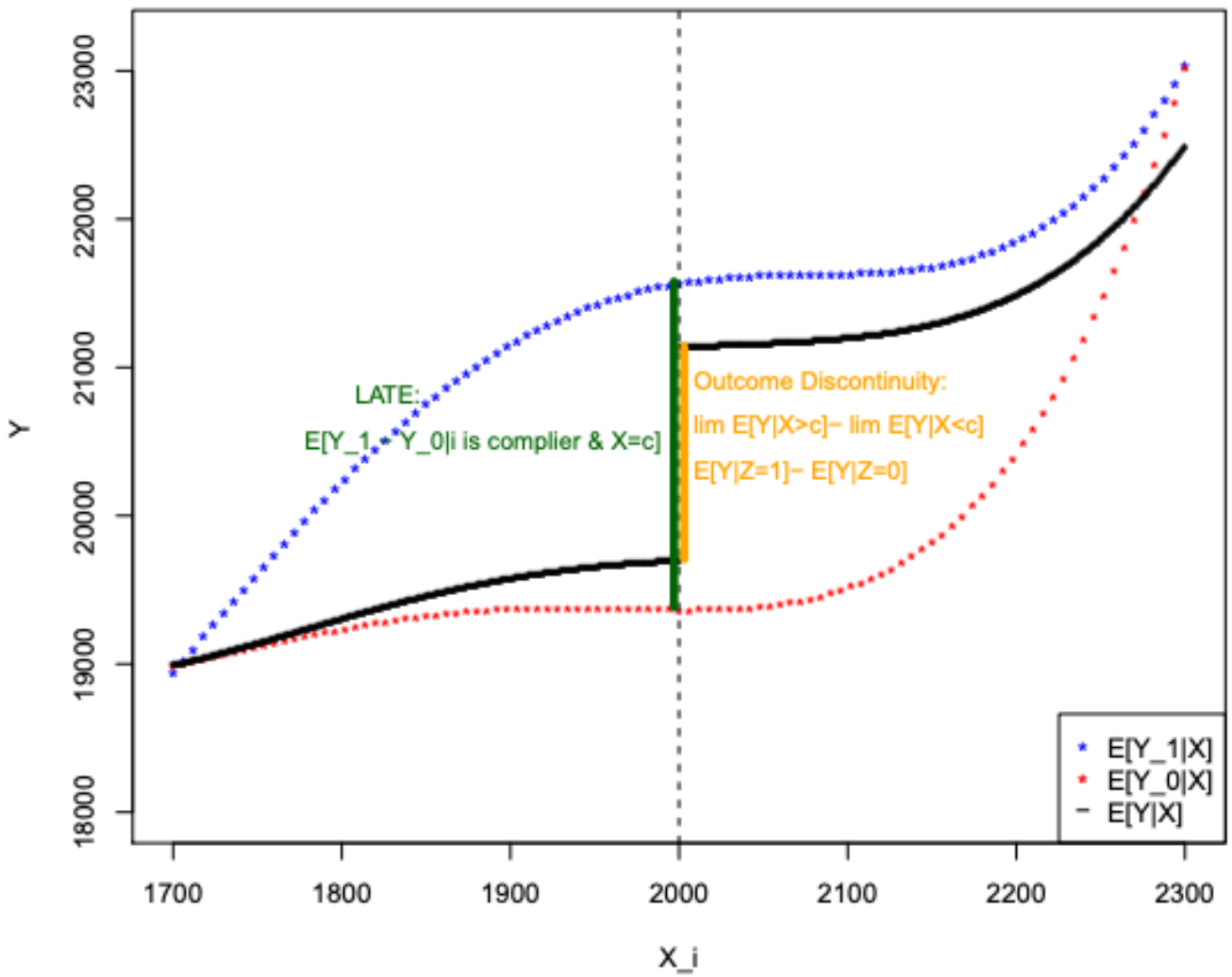
$$D_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \varepsilon_i$$



Second stage

- Estimating the outcome discontinuity using fitted values $\widehat{D}_i$ from the first stage:

$$Y_i = \beta_0 + \beta_1 X_i + \alpha \widehat{D}_i + u_i$$



Example 1

- **Question**: Do members of Party A curtail women's rights?
- **Running variable**:
  - Vote share– districts that are strongholds for Party A will be different in terms of their social, economic and cultural factors compared to strongholds of Party B.
- **Exogenous cutoff**:
  - Sharp, victory/loss (assuming a 2 party system, 50% vote share)
- **Bandwidth**:
  - 50%±4%
  - If we compare districts where candidates of that party narrowly won against those where they narrowly lost, those districts are very similar in terms of their **unobserved characteristics**

Example 2

- **Question**: Does an income-based cash transfer program decrease infant mortality?
- **Running variable**:
  - Income– those in extreme poverty face problems related to infant mortality that the rich do not (e.g. nutrition, access to healthcare)
- **Exogenous cutoff**:
  - Fuzzy, income-based assignment to treatment (e.g. need to make <£20k/yr)
- **Bandwidth**:
  - £20k±3k
  - If we compare people who narrowly qualify for a social program to those who narrowly fail to qualify, the latter may be similar enough to the former to act as a counterfactual.

Example 3

- **Question**: Does a test-based scholarship increase future earnings?
- **Running variable**:
  - Test score– those with really low test scores are different from those with high test scores in ways unrelated to the scholarship (e.g. wealth/ability to pay for tutoring, etc.)
- **Exogenous cutoff**:
  - Fuzzy, test score cutoff (e.g., need 90% to qualify for scholarship)
- **Bandwidth**:
  - 90%±5%
  - If we compare people who narrowly qualify for a the scholarship to those who narrowly fail to qualify, the latter may be similar enough to the former to act as a counterfactual.

Bringing it All Together

Kurdish Insurgency

- 44% of the population in Southeast living on less than US$1.1/day

- “The Kurdish identity question was expressed in terms of regional economic inequalities and suggested a socialist solution” (Yavuz, 2001: 10)
- Originally a “peasant movement”, recruiting largely from farming communities (Yavuz, 2001: 10)

Southeastern Anatolia Project (GAP)

- Infrastructural development project started in 1985
- Goals:
  - 19 dams
  - 22 hydropower plants
  - 1.8 million hectares of irrigation
- Income gains associated with switch to irrigated farming: 3-7x (Tokdemir et. al., 2016)
- Government hoped it would reduce appeal of PKK

TURKEY'S OUTREACH TO THE KURDS OF THE SOUTHEAST: GRAPPLING WITH THE ROOT CAUSES OF THE PKK PROBLEM

Date: 2008 January 31, 16:23 (Thursday) Canonical ID: 08ANKARA182_a

Original Classification: SECRET Current Classification: SECRET

WikiLeaks

is in general agreement with the government on the need for a broad approach, although with significant redlines. Most GOT officials believe greater focus on economic and social development in Turkey's southeast will have the greatest positive impact on ethnic Kurds' views towards the Turkish state. If Kurds are gainfully employed, have better educational opportunities, and see increased levels of infrastructure development throughout their region, their affinity for the terrorist PKK will wane further.

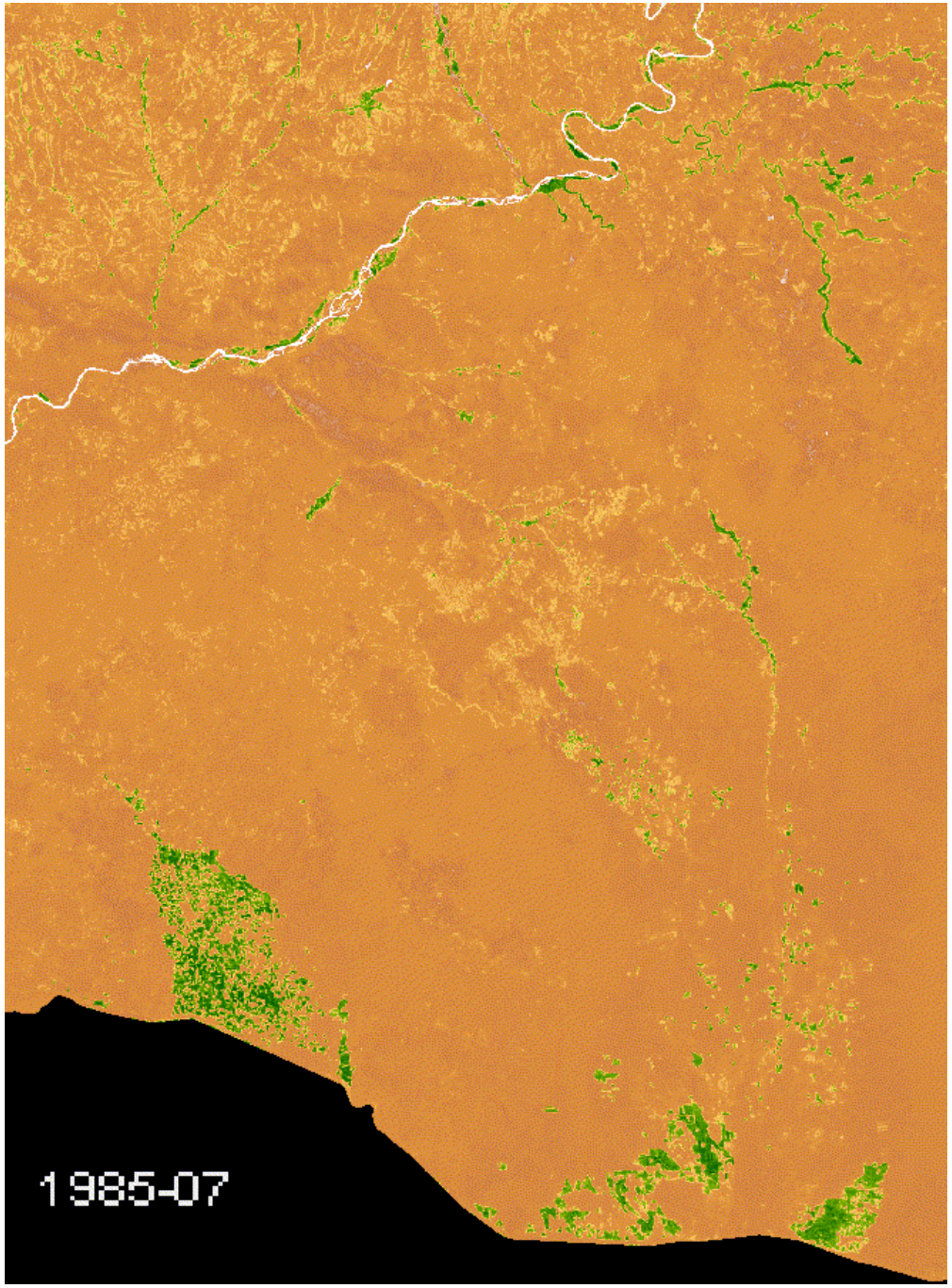
(Government Of Turkey)

4. (C) The GOT has ambitious plans, many of them laid out in the GOT's Action Plan. A top priority is ensuring completion of the massive Southeast Anatolia Project (GAP) within five years. Ekren told us he is also looking at creating a

Interviews with Farmers

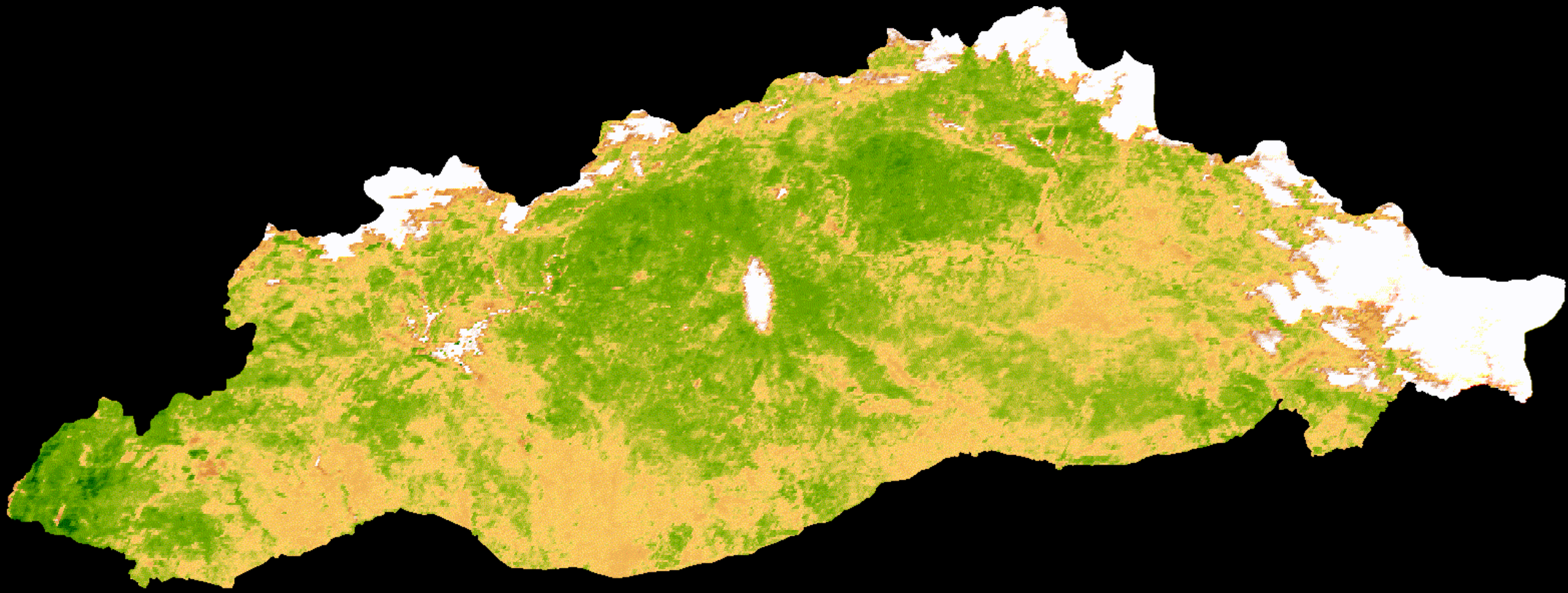
- Farmers who received GAP irrigation:
  - "**Our view of the state changed positively... we had hatred before, but now they started investing in the Southeast**"
  - "We did not trust the state before, but it brought electricity, water, phone, etc. to us and now we trust the state a lot"
  - "At last the state turned its face towards us"
  - "We did not see any accomplishments of governments in the past—we are a little bit happy to see some now." (Harris, 2016)

- Farmer who did not receive irrigation:
  - "Why should I not support Öcalan?"



Does irrigation reduce insurgent recruitment in Southeastern Turkey?

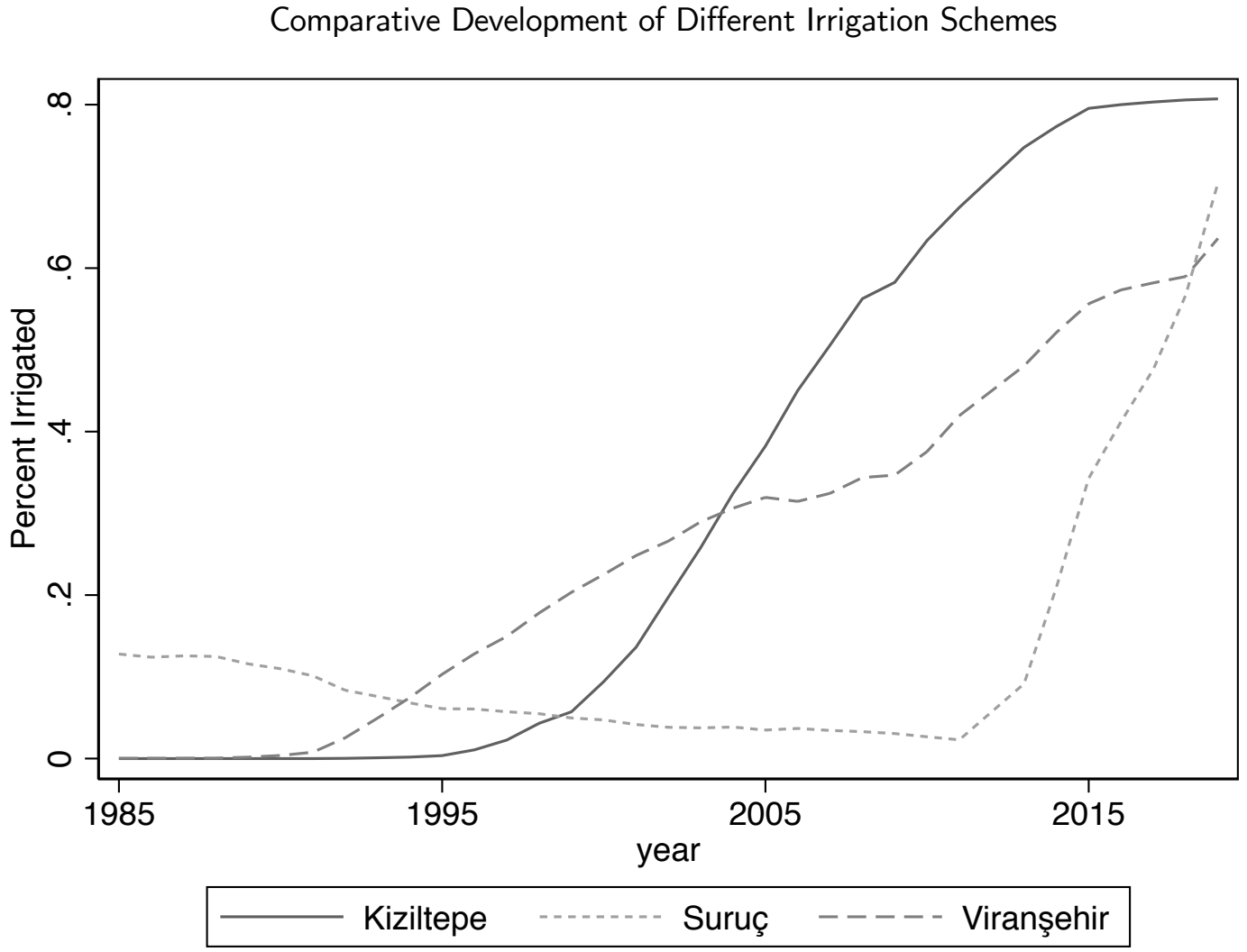
Irrigation



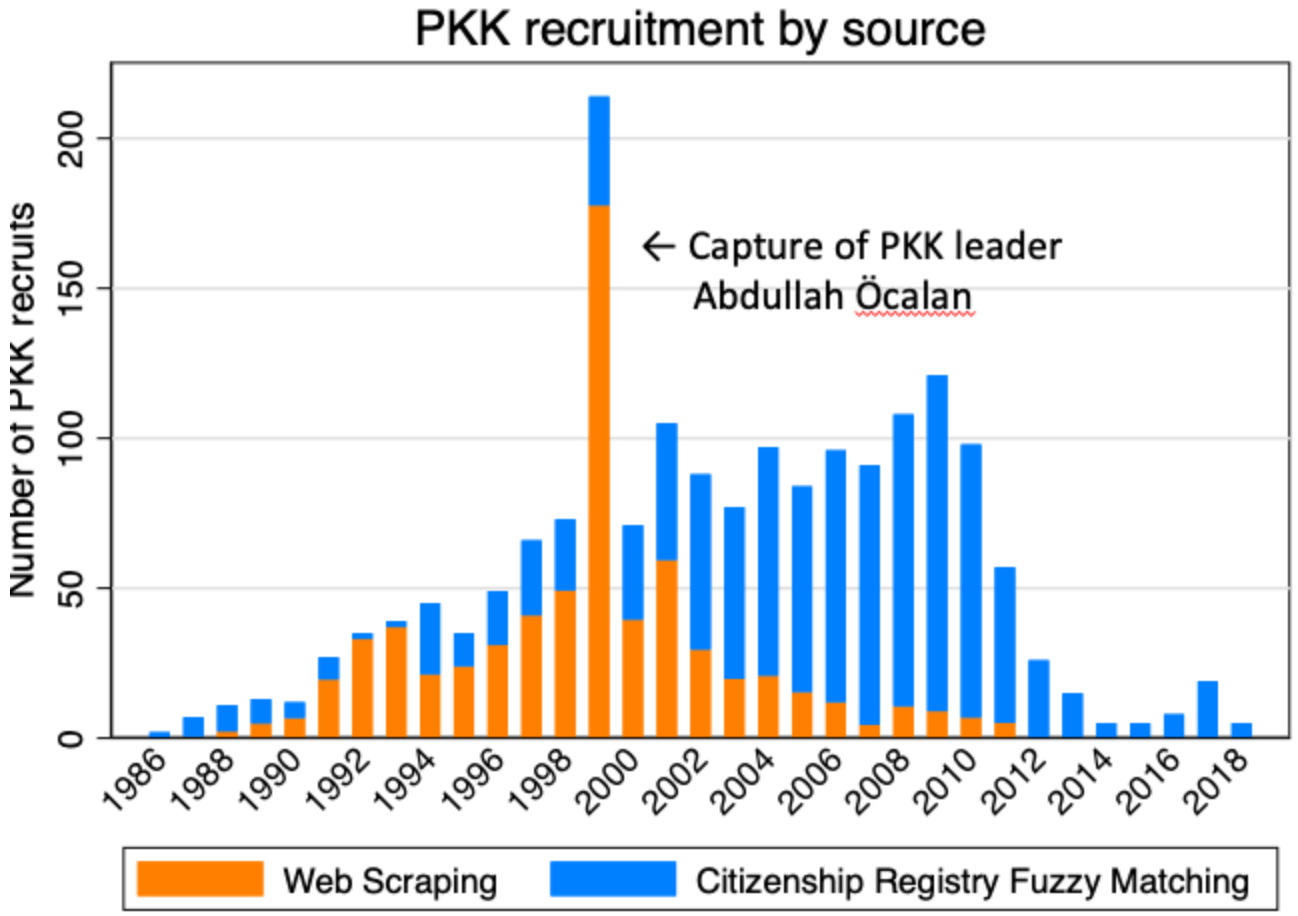
Data Source: LANDSAT 8

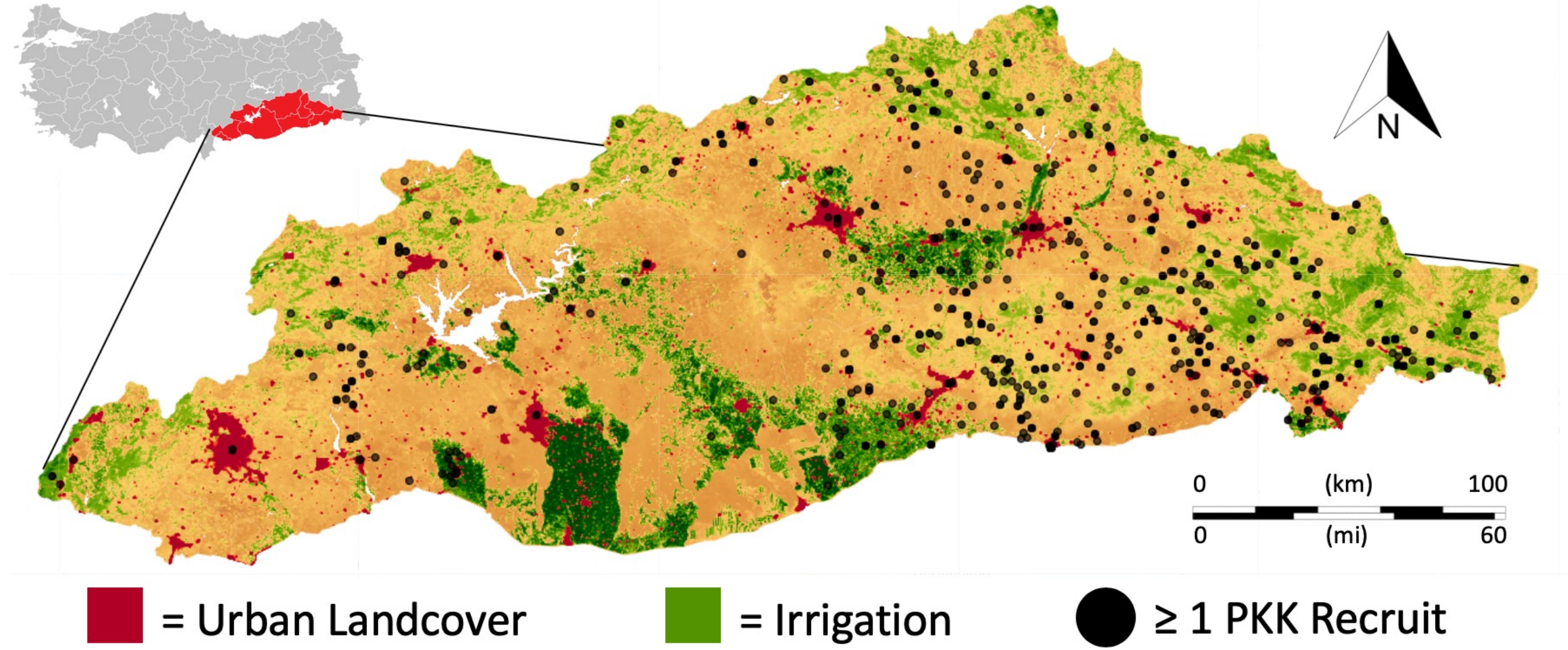
Irrigation

- Derived from satellite imagery analysis (NASA Landsat 5-8)
- Several different measures:
  - Percent of grid-cell under irrigation
  - Binary pre/post irrigation
  - Years since irrigation was introduced



Recruitment



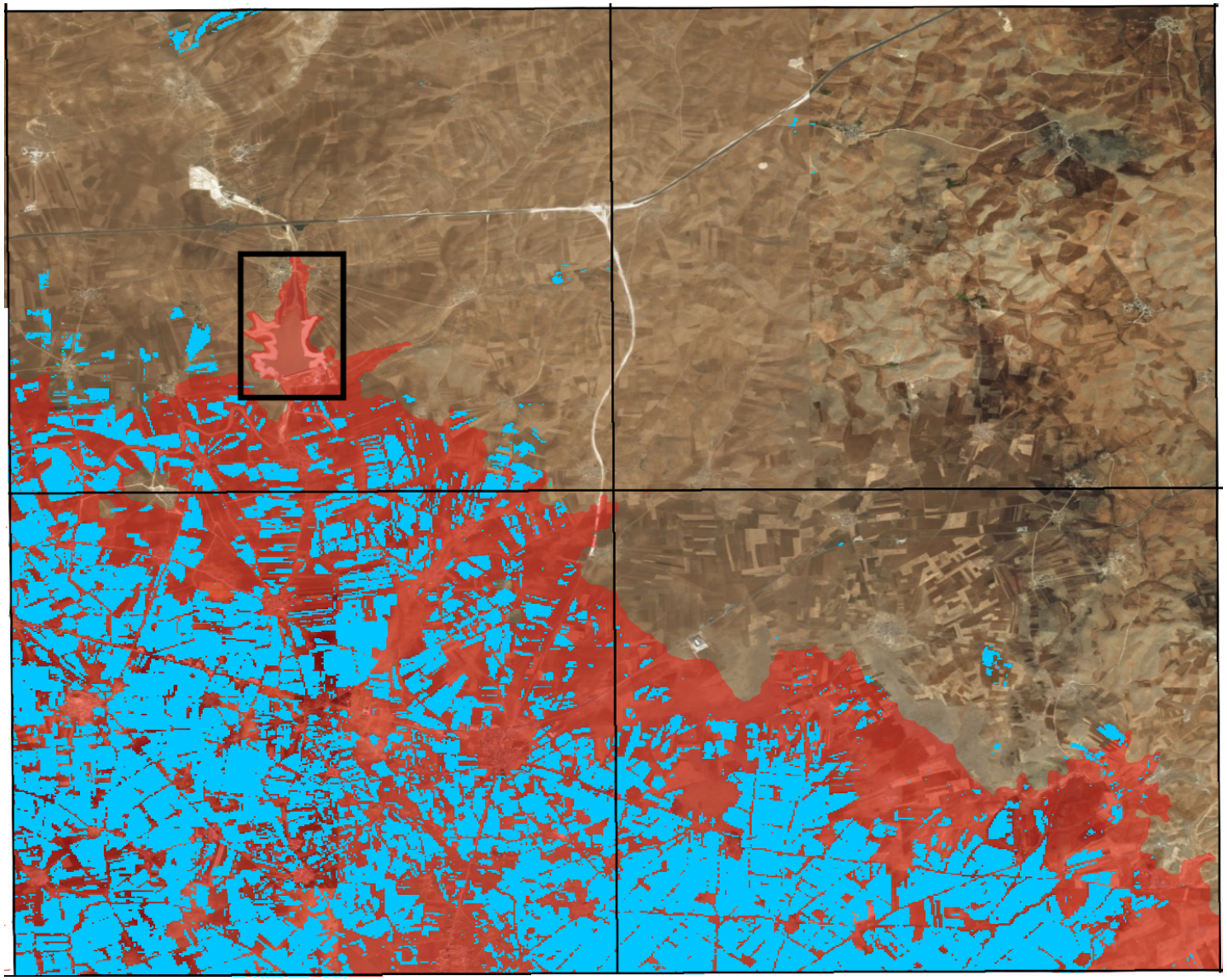

N
0
(km)
100
0
(mi)
60
= Urban Landcover
= Irrigation
≥ 1 PKK Recruit

RDD Framework

- **Question**: Does irrigation reduce recruitment?
- **Endogenous variable**:
  - Irrigation– a village in the mountains is very different from a village in the lowlands, in ways that probably affect PKK recruitment
- **Exogenous cutoff**:
  - Dam elevation
- **Bandwidth**:
  - If we compare *just* above and *just* below the altitude of a dam, they're probably going to be pretty similar in terms of demographics, politics, economy.
  - We could even make a plausible argument that the main difference between them is the ability to access irrigation

Treatment Variable

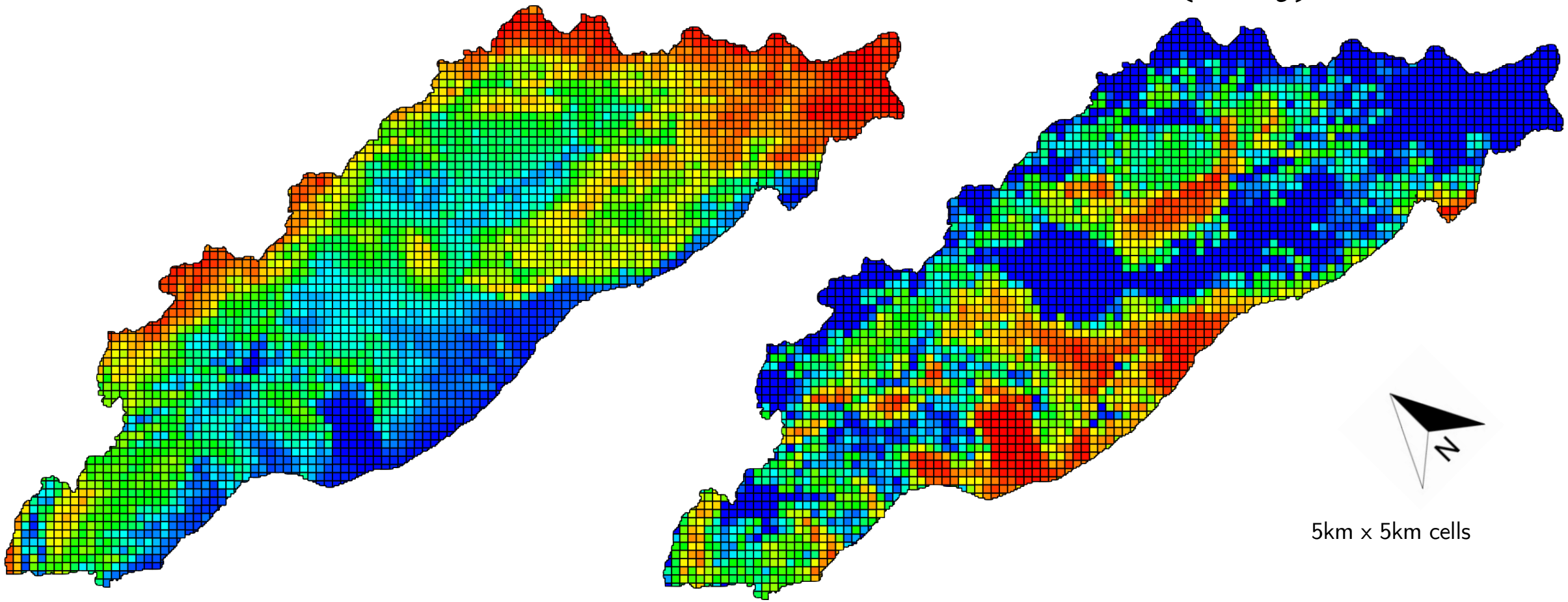
$$p(D_i = 1|X_i) = \begin{cases} f_1(X_i) \ \ if \ Z = 1 \\ 0 \ \ if \ Z = 0 \end{cases}$$



Terrain and Irrigation

$\ln(altitude_c \times (1 + slope_c))$

$\ln(IRR_c)$



5km x 5km cells

Figure 8: Discontinuity in PKK Recruitment Around Dam Altitude Cutoff

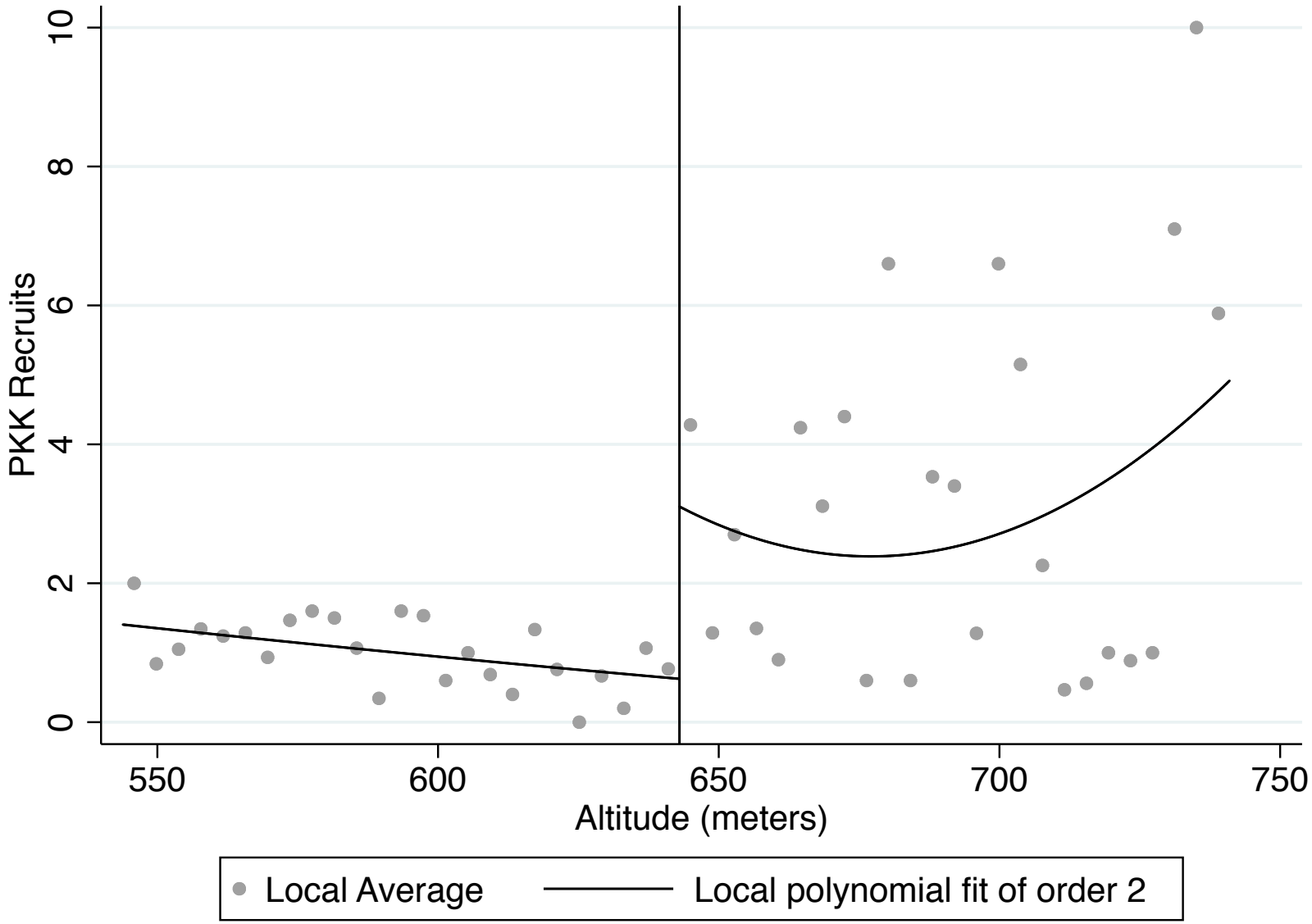


Table 6: Parametric FRD Results with Polynomial Manipulations

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 2SLS Regression						
Treatment Effect	-12.87**	-13.22**	-2.187	-11.15**	-11.83**	-1.835
	(5.311)	(5.902)	(3.636)	(4.602)	(5.270)	(3.322)
Panel B: First Stage						
Assigned Treatment	-.1836***	-.1744***	-.2483	-.1827***	-.1720***	-.2489***
	(.0516)	(.0594)	(.0676)	(.0499)	(.0597)	(.0664)
R-Squared	13.9%	14.2%	14.8%	21.5%	21.7%	22.5%
First stage F	12.63	8.601	13.47	13.2	8.29	14.05
p value	.000	.003	.000	.000	.004	.000
Covariates	No	No	No	Yes	Yes	Yes
Polynomial	Linear	Quadratic	Cubic	Linear	Quadratic	Cubic
Observations	1,945	1,945	1,945	1,868	1,868	1,868

$***p < 0.01, **p < 0.05, *p < 0.1$ This table reports estimates of the effect of being in a grid-cell just below the altitude of a feeder dam on PKK recruitment. Panel B reports the first stage results of the 2SLS specification, denoting the effect of the altitude cutoff on the probability of receiving irrigation. The first stage F statistic is reported below Panel B. Panel A reports the treatment effect derived from the second stage of a 2SLS regression. Columns 2 and 5 include a quadratic transformation of the running variable, and columns 3 and 6 contain a cubic transformation thereof. Columns 1, 2, and 3 contain no covariates, while columns 4, 5, and 6 contain all covariates. All specifications include the full battery of covariates and fixed effects.

Main Findings

- Irrigation is negatively related to conflict incidence and recruitment.
- Panel models indicate that a **27km^2 increase in irrigated area leads to a 49% drop in the probability of conflict.**
- Cross-sectional models suggest that **a fully irrigated 25km^2 area is 58% less likely than the average cell to experience a conflict event** involving Kurdish rebels during the 2016-2019 period

Any Questions?

o.ballinger@ucl.ac.uk