Canonical Research Designs VII: Regression Discontinuity III: Extensions

Paul Goldsmith-Pinkham

April 25, 2023

Roadmap for Today

- Last time: how to write an RD paper if everything works out smoothly
- This time: what are issues to keep in mind?
  - what can we do to account for hiccups?
- Also: what about Regression Kink design?

The asymptotic distribution of the RD estimator

- So far, we haven't discussed the asymptotic distribution of the RD estimator:

$$\tau_{SRD} = \lim_{z \downarrow 0} \hat{\mu}(z) - \lim_{z \uparrow 0} \hat{\mu}(z)$$

- A key discussion, however, was regarding the tradeoff between a large and small bandwidth on each side of $z = 0$
  - Small bandwidth – low bias, but very noisy
  - Big bandwidth – biased, but less noise
  - All asymptotic arguments need to be made "as the bandwidth shrinks"

- So, how fast should the bandwidth shrink?

The asymptotic distribution of the RD estimator

- A useful result from Cattaneo et al. (2020):

$$\text{MSE}(\hat{\tau}_{SRD}) = \text{Bias}^2(\hat{\tau}_{SRD}) + \text{Var}(\hat{\tau}_{SRD}) = (h^{2(p+1)}\mathcal{B})^2 + \frac{1}{nh}\mathcal{V}$$

  where p is the polynomial degree of the local linear estimator ($p = 1$ for linear), $\mathcal{B}$ is the leading bias term of an expansion, and $\mathcal{V}$ is the leading variance

- The choice of h that minimizes this MSE (conditional on p and the kernel) is

$$h_{MSE} = \left(\frac{\mathcal{V}}{2(p+1)\mathcal{B}^2}\right)^{1/(2p+3)} n^{-1/(2p+3)}$$

  e.g. for $p = 1$, $o(h_{MSE}) = n^{-1/5}$

The asymptotic distribution of the RD estimator

- But choosing $h_{MSE} \propto n^{-1/5}$ does not lead to zero bias in our distribution
- Remember the leading term for bias was $h^{2(p+1)} \to n^{-4/5}$, which is slower than n. E.g. we don't get the usual consistency necessary for our standard asymptotics.
  - Of course, this is all asymptotics... so it's all an approximation. But the thought experiment is important.
- One notable feature here: it is plausible the optimal bandwidth varies on each side of the cutoff.
  - If the variance of the outcome is higher on one side, it is plausible that the bandwidth would be wider to minimize MSE
  - This is easily accounted by allowing h^+ and h^-

The asymptotic distribution of the RD estimator

- The Calonico, Cattaneo and Titiunik (2015) approach advocates for using a plug-in estimator for bias over "undersmoothing"
  - What does this mean in practice?

- Rather than pick "small" bandwidths that are not MSE-optimal, try to account directly for bias by comparing a higher order estimate

- If you are using local linear regression, this means doing the following:
  - Estimate local *quadratic* estimate
  - Use the second derivative curvature to approximate the bias term for the local *linear* estimate, $\hat{\mathcal{B}}$
  - Use this estimator to adjust bias

- `rdrobust` does this automatically! Calonico, Cattaneo and Farrell (2018) show that this approach leads to correct inference, and preferable to undersmoothing

Let's look at it in practice

- BW Type: how is bandwidth chosen?
- Kernel: Choice of kernel (triangular is default, my preference is uniform but triangle has nice properties at edges)
- Effective # obs: how many obs within bandwidth?
- p: estimator polynomial order ($p = 1$ is linear)
- q: bias estimator polynomial order (must be at least 1 more than p)
- h is the bandwidth estimate for main estimate – b is the bandwidth for the bias estimator (needs more data usually, MSE optimal)
- Unique = mass points

```
Call: rdrobust

Number of Obs.                 2763
BW type                       mserd
Kernel                    Triangular
VCE method                        NN

Number of Obs.                 1376         1387
Eff. Number of Obs.             490          548
Order est. (p)                    1            1
Order bias  (q)                   2            2
BW est. (h)                   8.422        8.422
BW bias (b)                  13.990       13.990
rho (h/b)                     0.602        0.602
Unique Obs.                    1344         1313

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional     5.876     1.322     4.444     0.000     [3.284 , 8.468]
        Robust         -         -     3.606     0.000     [2.530 , 8.554]
=============================================================================
```

Let's look at it in practice

- Conventional: assumes zero bias in the inference method.
  - Coefficient and standard error are standard variance estimates that you get from running OLS with chosen *h*
  - No bias adjustment is done
- Robust: accounts for bias in two ways, but *only* in the inference
  - Centers CI around the *bias-adjusted* estimate
  - *Also* accounts for the additional noise from the estimation of the bias term
  - Can see the steps by using the `all` option

```
Call: rdrobust

Number of Obs.                 2763
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 1376         1387
Eff. Number of Obs.             490          548
Order est. (p)                    1            1
Order bias  (q)                   2            2
BW est. (h)                   8.422        8.422
BW bias (b)                  13.990       13.990
rho (h/b)                     0.602        0.602
Unique Obs.                    1344         1313

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional     5.876     1.322     4.444     0.000     [3.284 , 8.468]
        Robust         -         -     3.606     0.000     [2.530 , 8.554]
=============================================================================
```

Let's look at it in practice

- Conventional: assumes zero bias in the inference method.
  - Coefficient and standard error are standard variance estimates that you get from running OLS with chosen *h*
  - No bias adjustment is done
- Robust: accounts for bias in two ways, but *only* in the inference
  - Centers CI around the *bias-adjusted* estimate
  - *Also* accounts for the additional noise from the estimation of the bias term
  - Can see the steps by using the `all` option
- Cattaneo et al. (2020) advocate for:
  - Report $\hat{\tau}_{SRD}$ without bias adjustment (it is more MSE efficient than the bias corrected estimator)
  - Report the robust confidence interval

```
> summary(est)
Call: rdrobust

Number of Obs.                 2763
BW type                       mserd
Kernel                    Triangular
VCE method                        NN

Number of Obs.                 1376         1387
Eff. Number of Obs.             490          548
Order est. (p)                    1            1
Order bias  (q)                   2            2
BW est. (h)                   8.422        8.422
BW bias (b)                  13.990       13.990
rho (h/b)                     0.602        0.602
Unique Obs.                    1344         1313

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional     5.876     1.322     4.444     0.000     [3.284 , 8.468]
Bias-Corrected     5.542     1.322     4.191     0.000     [2.950 , 8.134]
        Robust     5.542     1.537     3.606     0.000     [2.530 , 8.554]
=============================================================================
```

What if we can't shrink our bandwidth? Discrete Regression Discontinuity

- Bias in the RD estimates comes from the approximation of the conditional mean function
  - The smaller the bandwidth, the better the local approximation!
- What if the running variable is discrete? E.g., age
- Kolesar and Rothe (2018) discuss exactly this scenario and propose an "Honest" RD estimation approach which approximates the bias by asssuming a maximum second derivative

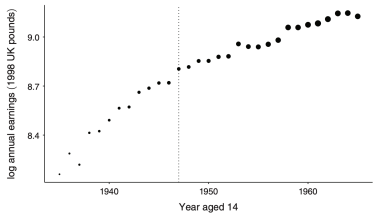


FIGURE 2. AVERAGE OF NATURAL LOGARITHM OF ANNUAL EARNINGS BY YEAR AGED 14

Notes: Vertical line indicates the year 1947, in which the minimum school-leaving age changed from 14 to 15. Volume of dots is proportional to share of workers in the full data with the corresponding age.

What if we can't shrink our bandwidth? Discrete Regression Discontinuity

- Kolesar and Rothe say – ignore the bias-adjustment fact, and just assume undersmoothing. If this is the case, you can just use standard Eicker-Huber-White errors (e.g. "robust" s.e.)
  - But we know this argument only works if you can get enough observations "close" to the cutoff
  - With discrete variables, this fails

- Recall that we're extrapolating the conditional mean function to the cutoff
  - If we are willing to put a bound on the 2nd derivative function, and assume it is in a class of Hölder functions (which is very general), we can bound our maximum bias, and use this adjust our confidence intervals

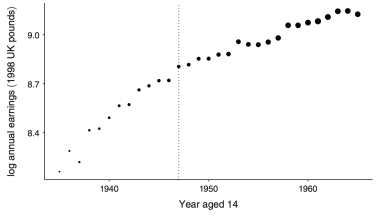


FIGURE 2. AVERAGE OF NATURAL LOGARITHM OF ANNUAL EARNINGS BY YEAR AGED 14

Notes: Vertical line indicates the year 1947, in which the minimum school-leaving age changed from 14 to 15. Volume of dots is proportional to share of workers in the full data with the corresponding age.

What if we can't shrink our bandwidth? Discrete Regression Discontinuity

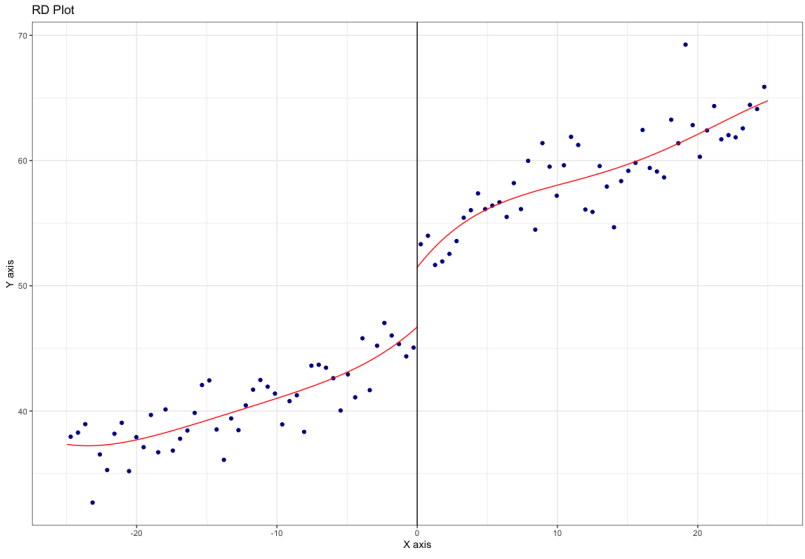
- This approach is valid even with continuous running variables, and is a powerful and robust way to allow for misspecification bias

- My experience has been that this package (`RDHonest`) works much better with discrete RV than `rdrobust` (which has adjustments for discreteness)

- However, you need to choose your maximum bias, and it can be challenging to do so *ex ante*.
  - Armstrong and Kolesar discuss ways to do this in a data driven way given additional assumptions.
  - You should also show how your results change with this parameter choice, similar to bandwidth!

Comparing RDHonest and rdrobust

- First, let's look at the plot
- RDRobust estimate is 5.9 (2.5,8.5)
- RDHonest estimate varies depending on choice of second derivative bound. For $M = 0.1$: 5.9 (2.9, 8.8)
- What would you pick for M?



Comparing RDHonest and rdrobust

- First, let's look at the plot
- RDRobust estimate is 5.9 (2.5,8.5)
- RDHonest estimate varies depending on choice of second derivative bound. For $M = 0.1$: 5.9 (2.9, 8.8)
- What would you pick for M?

```
> summary(est)
Call: rdrobust

Number of Obs.                 2763
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 1376         1387
Eff. Number of Obs.             490          548
Order est. (p)                    1            1
Order bias  (q)                   2            2
BW est. (h)                   8.422        8.422
BW bias (b)                  13.990       13.990
rho (h/b)                     0.602        0.602
Unique Obs.                    1344         1313

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional     5.876     1.322     4.444     0.000     [3.284 , 8.468]
        Robust         -         -     3.606     0.000     [2.530 , 8.554]
=============================================================================
```

Comparing RDHonest and rdrobust

- First, let's look at the plot
- RDRobust estimate is 5.9 (2.5,8.5)
- RDHonest estimate varies depending on choice of second derivative bound. For $M = 0.1$: 5.9 (2.9, 8.8)
- What would you pick for M?

```
> est_rdhonest
Call:
RDHonest(formula = lee08$voteshare ~ lee08$margin, cutoff = 0,     M = 0.1, opt.criterion = "MSE")


Inference by se.method:
   Estimate Maximum Bias Std. Error
nn 5.901309    0.7851139   1.309639

Confidence intervals:
nn     (2.935053, 8.867566), (2.962031, Inf), (-Inf, 8.840588)

Bandwidth: 8.603387
Number of effective observations: 206.84
```

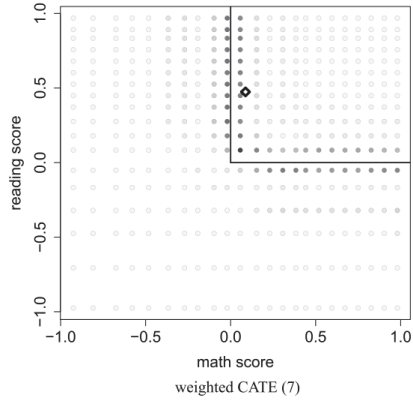
Comparing RDHonest and rdrobust

- First, let's look at the plot
- RDRobust estimate is 5.9 (2.5,8.5)
- RDHonest estimate varies depending on choice of second derivative bound. For $M = 0.1$: 5.9 (2.9, 8.8)
- What would you pick for M?

	mu	lb	ub	case	bw
	<dbl>	<dbl>	<dbl>	<chr>	<dbl>
1	5.88	2.95	8.13	RDRobust	8.42
2	7.51	5.55	9.47	RDHonest M=0.01	21.9
3	5.90	2.96	8.84	RDHonest M=.1	8.60
4	8.78	4.58	13.0	RDHonest M=1	3.49

What if we can't shrink our bandwidth? Discrete Regression Discontinuity

- Another similar paper in this space is Imbens and Wager (2019)
  - The estimation approach also puts bound on the second derivative, but uses a numerical approach to do estimation
  - Package in R: `optrdd`
- This approach potentially has slightly smaller confidence intervals
- It also generalizes to multivariate RD settings (e.g. spatial settings) quite easily
  - See also `rdmulti`, which also allows for multiple cutoffs



Multiple Cutoffs

- Cattaneo, Titiunik, Vazquez-Bare and Keele (2016) and Bertanha (2020) touch on what to do when not every threshold is the same (See the discussion in Cattaneo, Idrobo and Titiunik's extensions textbook for a very clean discussion)
  - E.g. **MultiThreshold** contrast a political election with 2 vs. 3 candidates – what is the "winning" threshold?
  - E.g. **MultiCutoff** Cutoff based on two test scores
- These have separate complications associated with them

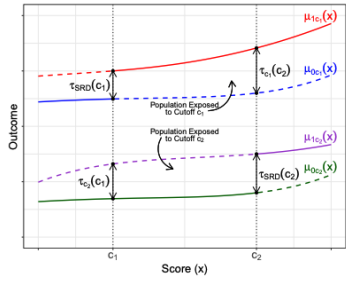
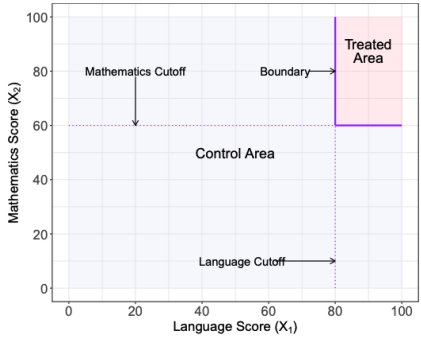


Figure 5.2: Multi-Cutoff RD Design with Two Non-cumulative Cutoffs

Multiple Cutoffs

- Cattaneo, Titiunik, Vazquez-Bare and Keele (2016) and Bertanha (2020) touch on what to do when not every threshold is the same (See the discussion in Cattaneo, Idrobo and Titiunik's extensions textbook for a very clean discussion)
  - E.g. **MultiThreshold** contrast a political election with 2 vs. 3 candidates – what is the "winning" threshold?
  - E.g. **MultiCutoff** Cutoff based on two test scores

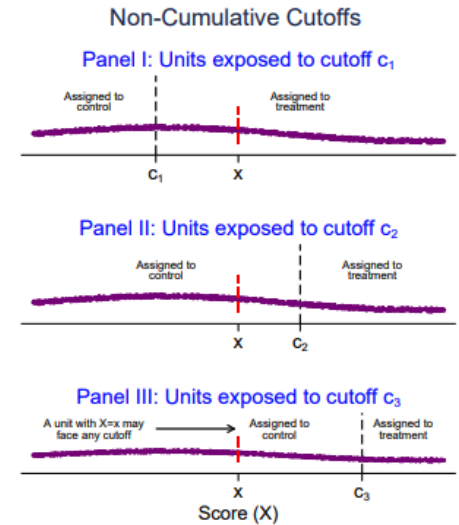
- These have separate complications associated with them



(a) Hard-threshold Assignment

Multiple Threshold - Cumulative vs. Not

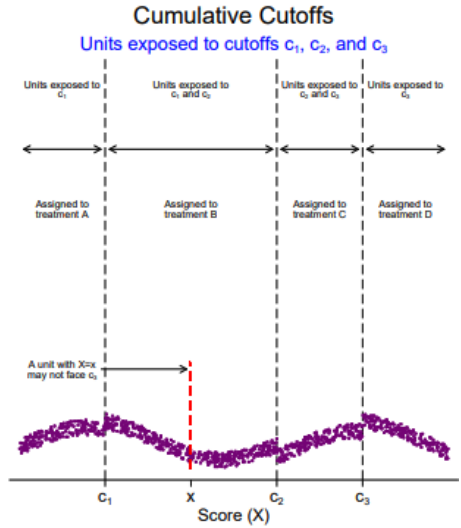
- Do the treatment cutoffs vary based on unit, or are there just many of them?
- An example non-cumulative case is when the thresholds vary on geography: all units could potentially be exposed to treatments defined at different cutoffs
- An example cumulative case is where dosing is a function of the score – those facing different treatment cutoffs are not comparable (e.g. a support problem in the running variable)



(a) Non-cumulative Cutoffs

Multiple Threshold - Cumulative vs. Not

- Do the treatment cutoffs vary based on unit, or are there just many of them?
- An example non-cumulative case is when the thresholds vary on geography: all units could potentially be exposed to treatments defined at different cutoffs
- An example cumulative case is where dosing is a function of the score – those facing different treatment cutoffs are not comparable (e.g. a support problem in the running variable)



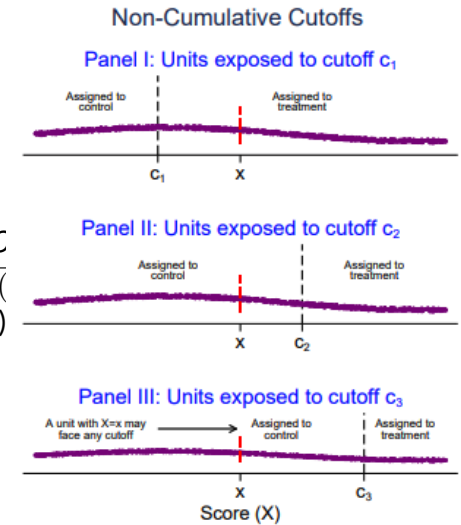
(b) Cumulative Cutoffs

Multiple Threshold - Non-cumulative pooling

- With non-cumulative treatments, can consider a cutoff specific treatment and proceed identically – each one identified separately

- To pool these estiamtes, you can normalize and center at the same point

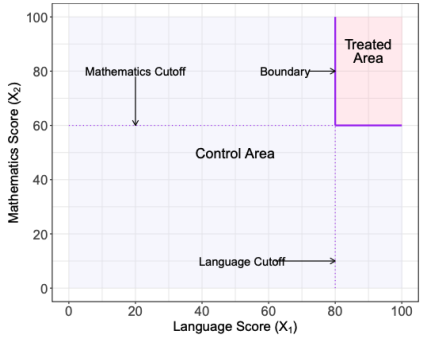
$$\tau_{SRD}^{pool} = \sum_c w(c)\tau_{SRD}(c), \; w(c) = \frac{f_{X|C}(c|c)Pr(C}{\sum_c f_{X|C}(c|c)Pr(} \quad (1)$$



(a) Non-cumulative Cutoffs

Multiple Cutoffs

- With multiple scores, the limits are defined in a multi-dimensional way.
- Integrating over these points has a similar logic to multiple-threshold



(a) Hard-threshold Assignment

What about bunching? Bounds on Treatment effects

- Recall our Mcrary (2008) test for bunching in the running variable
  - Concern is maniuplated running variable
- For example, in Dee et al. (2019), there is clear manipulation, and so using this RD directly would be spurious
- This is a particularly egregious case – many are less stark
  - Moreover, you may also be underpowered to detect the break if they do exist!

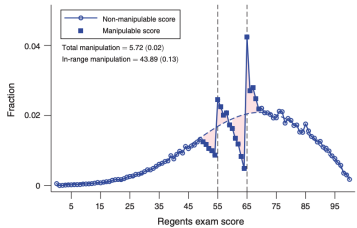


FIGURE 1. TEST SCORE DISTRIBUTIONS FOR CORE REGENTS EXAMS, 2004–2010

Notes: This figure shows the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004–2010. Core exams include English Language Arts, Global History, US History, Math A/Integrated Algebra, and Living Environment. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points indicating a manipulable score. The dotted line beneath the empirical distribution is a subject-by-year specific sixth-degree polynomial fitted to the empirical distribution excluding the manipulable scores near each cutoff. The shaded area represents either the missing or excess mass for manipulable scores as we define based on the scoring guidelines described in Section III and detailed in online Appendix Table A3. Total manipulation is the fraction of test takers with manipulated scores. In-range manipulation is the fraction of test takers with manipulated scores normalized by the average height of the counterfactual distribution to the left of each cutoff. Standard errors are calculated using the parametric bootstrap procedure described in the text. See the online Data Appendix for additional details on the sample and variable definitions.

What about bunching? Bounds on Treatment effects

- Gerard, Rokkanen and Rothe (2020) propose a partial identification approach to allow for the possibility of bunching
- Think of this as a check on the robustness of the results – how sensitive are the results to manipulation?
  - Code is available for R and Stata: `rdbounds`
- The key result hinges on the idea that the manipulation goes in one direction, and that these right-ward manipulated individuals "mask" the true underlying effect
  - These individuals are always manipulated to the right side of the cutoff (they are the "excess mass")
- Derive sharp bounds for those individuals who can actually be affected by the treatment
  - Do this by identifying the share of the "masking" individuals: $\tau = 1 - f_Z(0^-)/f_Z(0^+)$.
  - If Mcrary test holds, this number is zero!
- This is a very nice approach if you have bunching issues in your design

An important question

- Imagine you have an RD that you ran, and you want to compare across estimates across two groups ($W_i \in (0, 1)$)
  - How would you do this?

- Note how you would do it in a simple OLS setting:

$$y_i = \alpha_0 + Z_i\gamma_0 + Z_i 1(Z_i > 0)\delta_0 + 1(Z_i > 0)\tau_{SRD} + W_i\alpha_1 + W_i Z_i\gamma_0 + W_i Z_i 1(Z_i > 0)\delta_0 + W_i 1(Z_i > 0)\tau_{SRD,diff} + \epsilon_i$$

  $\tau_{SRD,diff}$ would give you the difference – this approach is easy if you pick a bandwidth and use a uniform kernel

- But now we have all these better tools (and we need them in a lot of places) – how could still do this?
  - Apply the delta method to the transformation we want to do!

Comparing coefficients across setups

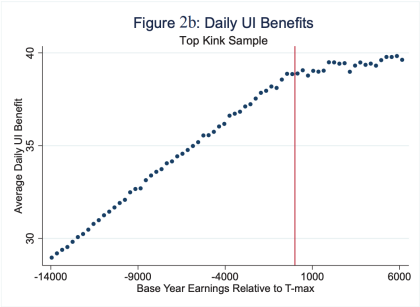
- Consider an RDRobust estimate for $\tau(W = 1)$ and $\tau(W = 0)$.
  - The difference is $f(\tau_1, \tau_0) = \tau_1 - \tau_0$

- Recall that our variance estimate is $(f')^T \Sigma f'$, where f' is the gradient vector and Σ is our variance covariance matrix of τ_1 and τ_0.
  - How do we get Σ? Well, recall that we estimated these separately from one another, so the covariance terms are zero
  - Hence, Σ is just the diag($Var(\tau_1)$, $Var(\tau_0)$)

- In our simple example, $Var(\tau_1 - \tau_0) = Var(\tau_1) + Var(\tau_0)$ and so Delta method is exact
  - However, can study more complicated functions as well!
  - This approach also works with the `RDHonest` approach as well, just need to account for the additional bias terms (see Appendix of Armstrong and Kolesar (2020))

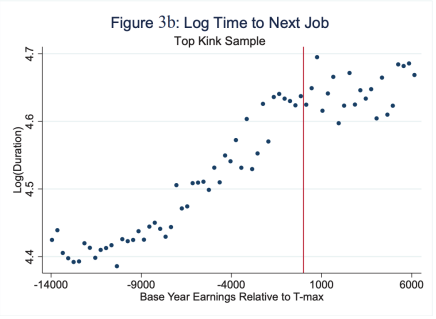
Regression Kink Design

- Much of our discussion centered around a discontinuous jump in the outcome (and treatment variable)
- Nielsen et al. (2010) initially coin the concept instead of a regression kink design (further worked on by Card, Lee, Pei and Weber (2016)
  - Key difference here exploits a difference in *slope*, rather than a level difference
  - Technically, could be both!
- This approach is very powerful because many policy tools have linear shifts in incentives, rather than jumps



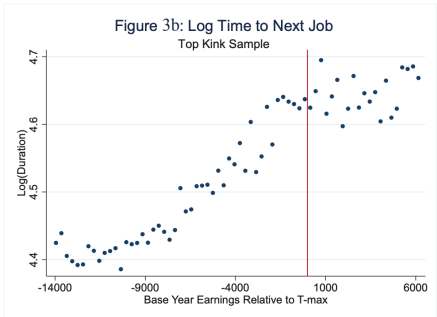
Regression Kink Design

- Much of our discussion centered around a discontinuous jump in the outcome (and treatment variable)
- Nielsen et al. (2010) initially coin the concept instead of a regression kink design (further worked on by Card, Lee, Pei and Weber (2016)
  - Key difference here exploits a difference in *slope*, rather than a level difference
  - Technically, could be both!
- This approach is very powerful because many policy tools have linear shifts in incentives, rather than jumps



Regression Kink Design

- This is a deeply interesting approach, but requires a *lot* of data
- Why? because slope changes are hard to see. Did the slope change because of curvature, of because of a kink?
- Ganong and Jäger (2018) discuss a test for this to account for this fact
- This is a good approach to keep in mind! See, e.g., Indarte (2021)



Regression Kink Design

- This is a deeply interesting approach, but requires a *lot* of data
- Why? because slope changes are hard to see. Did the slope change because of curvature, of because of a kink?
- Ganong and Jäger (2018) discuss a test for this to account for this fact
- This is a good approach to keep in mind! See, e.g., Indarte (2021)

Figure 2: The Effect of Seizable Equity on Bankruptcy Filings