Causal Inference II

Mixtape Session

Roadmap

Continuous DiD

- Dose causal parameter
- Identification
- Selection bias
- Interpreting TWFE

Concluding remarks

DiD Revolution

Standard DiD model with binary treatment D_{it} estimated with TWFE

$$Y_{it} = \delta D_{it} + \alpha_i + \tau_t + \varepsilon_{it}$$

With differential timing, TWFE estimates of δ are biased when treatment effects "change over time" (dynamic), even when $E[\Delta Y^0 | D = 1] = E[\Delta Y^0 | D = 0]$.

Borusyak, et al. (2022), de Chaisemartin and D'Haultfouielle (2020), Sun and Abraham (2020), Goodman-Bacon (2021)

Alternative methods

- OLS loves to exploit variation, but this can lead to naive causal inference practices
- Causal inference is not focused on exploiting variation but rather estimating treatment effects using appropriate counterfactuals
- Alternative estimators avoided “forbidden contrasts” altogether allowing us to entertain unrestricted heterogeneity and obtain reasonable versions of ATT parameter

Continuous DiD

- A very common panel model will use a treatment variable which is continuous, not binary
- Examples include minimum wage papers, my JHR on abortion clinic closures causing increased travel distance, vaccinations, price elasticity of demand etc.
- Variation is in “treatment intensity” and researchers typically use TWFE for estimation, or perhaps count models like Poisson

Quotes

"The two-period regression estimator can be easily modified to allow for continuous, or at least non-binary, treatments." (Wooldridge 2005)

"A second advantage of regression DiD is that it facilitates the study of policies other than those that can be described by a dummy." (Angrist and Pischke 2008)

Continuous DiD

Hani Mansour @hnmansour · Dec 14, 2020
I remember seeing a paper about estimating an event study with a continuous variable but can't seem to track it. Any leads #EconTwitter @causalinf @agoodmanbacon?
7 8 27

David Burgherr @D_Burgherr · Mar 25, 2020
On this point, I am very curious how much the issues you point out with DD designs -- variance-weighting of treatment effects (TE) and bias in the case of time-varying TE -- matter for continuous treatment variables. Do you have any take or reference on that?

Thanks in advance!
2

Khoa Vu @KhoaVuUmn · Nov 16, 2020
#EconTwitter Question on DiD: I'm looking for a reference on what to do when the treatment variable is continuous and you suspect that the effect is nonlinear, e.g. medium exposure might have bigger effect than high exposure.
4 7 42

Ben Glasner @BenGlasner · Oct 29, 2020
Any recs on DID packages in r for multiple treatment periods with different timings and continuous treatment values? Think minimum wage changes over time? Something to compare TWFE against... #econtwitter
2

Michelle Spiegel @michspieg · Apr 22, 2020
I am writing a DiD paper with a continuous treatment. Any paper recommendations to help think about statistical power in this context? #EconTwitter #soctwitter #AcademicTwitter

Kait Sims @kaitmsims · Aug 25, 2020
#EconTwitter recommendations for event study/DiD papers with staggered treatment time, continuous treatment intensity, and where treatment can turn on and off more than once for the same individual?

3 4 4

Jason Baron @JasonBaron4 · Apr 21, 2020
I know there have been previous threads on the most recent DID papers, but does anyone know if there are any recent methodological papers specifically looking at DID implementation with a continuous treatment variable? @causalinf @jondr44
4 9 37

Adam Roberts @adamn_roberts · Mar 28
Question for DiD experts:
Is there a heterogenous treatment effects solution that works for continuous treatments? I'm specifically thinking about early childhood intervention papers that define treatment as "age 0-5" exposure to something like county food stamps availability.
1 7

Adam Roberts @adamn_roberts · Mar 28
This type of treatment has staggered timing and enough heterogeneity to make TWFE a poor approach but after diving into the new DiD literature I'm struggling to figure out the "correct" approach with a continuous treatment variable. Any thoughts @causalinf @Andrew__Baker ?
4

Nicholas Reynolds @nick_reynolds88 · Apr 28
Does anyone know of papers deriving what TWFE with continuous treatment and allowing for heterogenous treatment effects estimates?

My intuition is same "problems" found for staggered diff-in-diff would exist here ... but notation would explode so no one has written it out?
1 2 2

Peter Bergman @peterbergman_ · May 11, 2020
Seems like "dosage"/"intensity" diff-in-diff--where there aren't 2 groups but a continuous measure w/ varying intensity of treatment--requires potentially stronger identifying assumptions than DID for 2 groups. Is this discussed in any of the recent DID lit updates? cc @causalinf
6 5 47

Nick Hagerty @hagertynw · Mar 29
Conceptually it's not that distinct right? We're still trying to identify off similar shocks in different places at different times. I thought the main difference is that our variables are continuous treatments -- algebra is harder but papers prob. coming in next couple years
3 6

Michael Wiebe @michael_wiebe · Feb 9
Who's writing the @agoodmanbacon paper on diff-in-diff with a continuous treatment variable (instead of binary)?

#EconTwitter
1

Davide Proserpio @dade_us · Apr 12, 2020
Looking for recommendations about DD papers where the treatment is continuous, thanks! #EconTwitter
6 1 8

Overview

1. What of what we have learned carries forward to the continuous case?
2. Some of the problems with continuous (maybe most) don't even have to do with differential timing, so I'm not going to cover it

Introducing a new causal parameter

- **ATT**: Extensive margin causal parameter. Do this versus don't do this.
- **Dose**: Intensive margin causal parameter. Do this much versus this much.

The dose causal parameter will be based on Angrist and Imbens (1995)

Parameters

Average treated on the treated

$$ATT(d|d) = E[Y_{it}^{d} - Y_{it}^{0}|D_{it} = d]$$

while the treatment, D, can be any amount, d, that amount is technically a particular dose. We raised the minimum wage, but we raised it to a particular wage.

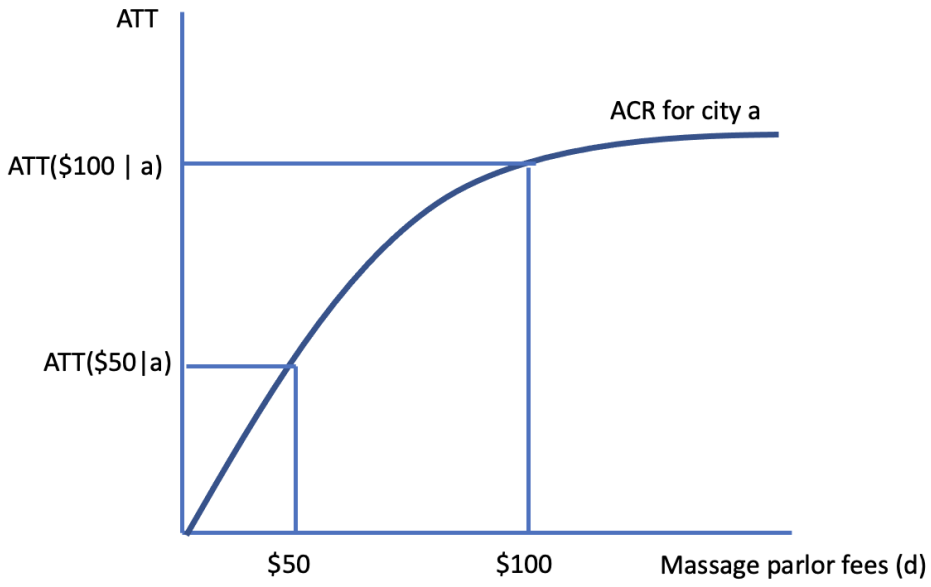
Parameters

Average treated on the treated

$$ATT(d|d) = E[Y_{it}^{d} - Y_{it}^{0}|D_{it} = d]$$

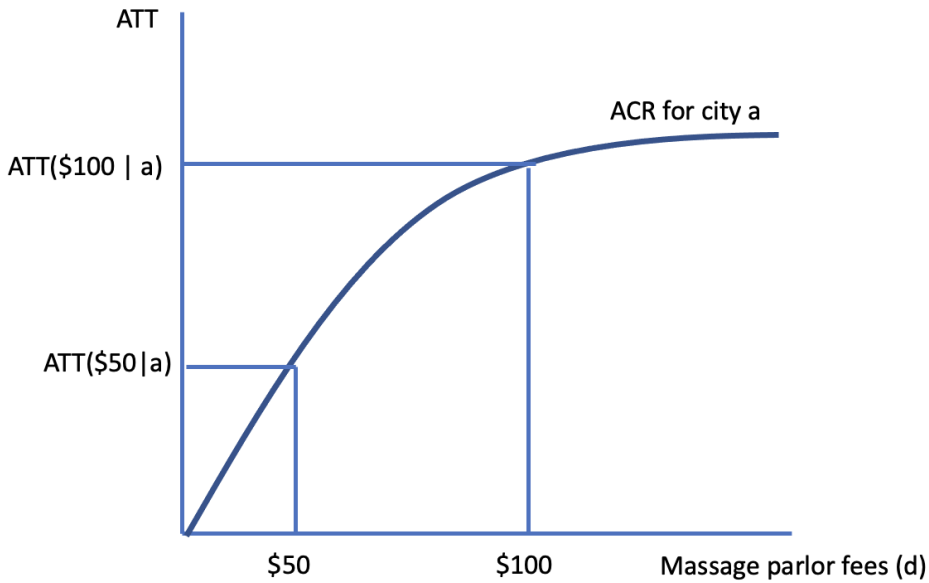
This is “the ATT of d for the groups that chose d dosage” which uses as its comparison no dose.

ATT for a given dose



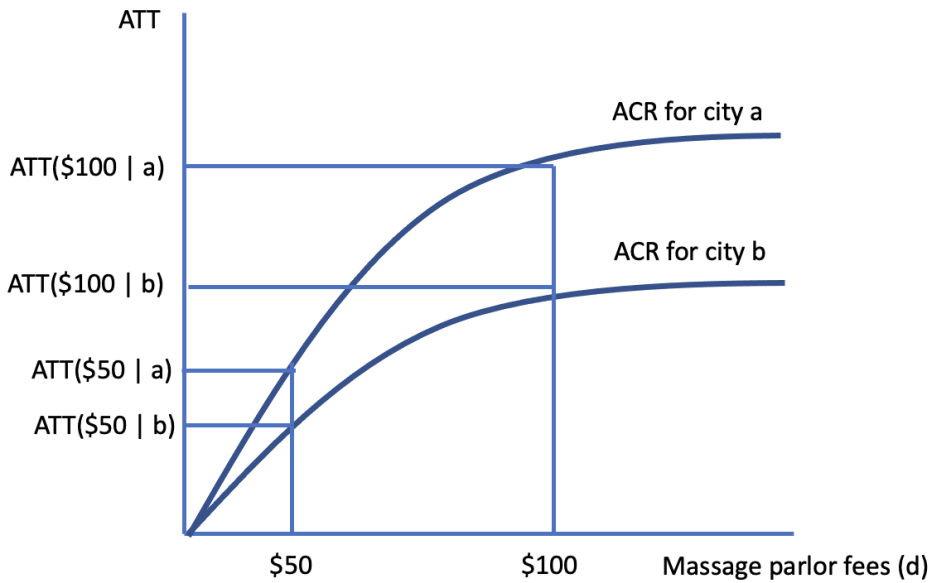
What is the effect of setting fees to $100 versus nothing at all? It's ATT($100—a) for this city.

ATT for a given dose



Assume city a did choose $d = \$100$. Then ATT($50—a) just means that that is its ATT *had* it chosen the lower level. The curve, in other words, is tracing out all average causal response for this city.

ATT for a given dose



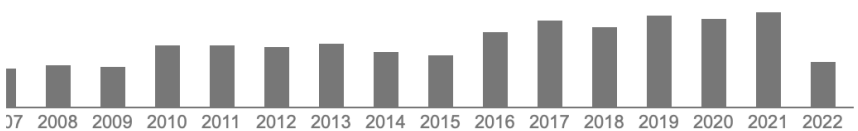
What if everyone has different responses? In other words, city a has the higher curve than city b. Then there are several comparisons possible. What is the effect of $50 on outcomes for cities that actually chose $50 versus those than actually chose $100?

Average causal response function

guido imbens

Two-stage least squares estimation of average causal effects in models with variable treatment intensity

Authors	Joshua D Angrist, Guido W Imbens
Publication date	1995/6/1
Journal	Journal of the American statistical Association
Volume	90
Issue	430
Pages	431-442
Publisher	Taylor & Francis Group
Description	Two-stage least squares (TSLS) is widely used in econometrics to estimate parameters in systems of linear simultaneous equations and to solve problems of omitted-variables bias in single-equation estimation. We show here that TSLS can also be used to estimate the average causal effect of variable treatments such as drug dosage, hours of exam preparation, cigarette smoking, and years of schooling. The average causal effect in which we are interested is a conditional expectation of the difference between the outcomes of the treated and what these outcomes would have been in the absence of treatment. Given mild regularity assumptions, the probability limit of TSLS is a weighted average of per-unit average causal effects along the length of an appropriately defined causal response function. The weighting function is illustrated in an empirical example based on the relationship between schooling and earnings.
Total citations	Cited by 1372



Scholar articles

Two-stage least squares estimation of average causal effects in models with variable treatment intensity
JD Angrist, GW Imbens - Journal of the American statistical Association, 1995
Cited by 1358 Related articles All 14 versions

Average causal response with variable treatment intensity *
J Angrist, G Imbens - 1995
Cited by 16 Related articles All 10 versions

Angrist and Imbens 1995

*"We refer to the parameter β as the **average causal response (ACR)**. This parameter captures a weighed average causal responses to a unit change in treatment, for those whose treatment status is affected by the instrument. …"*

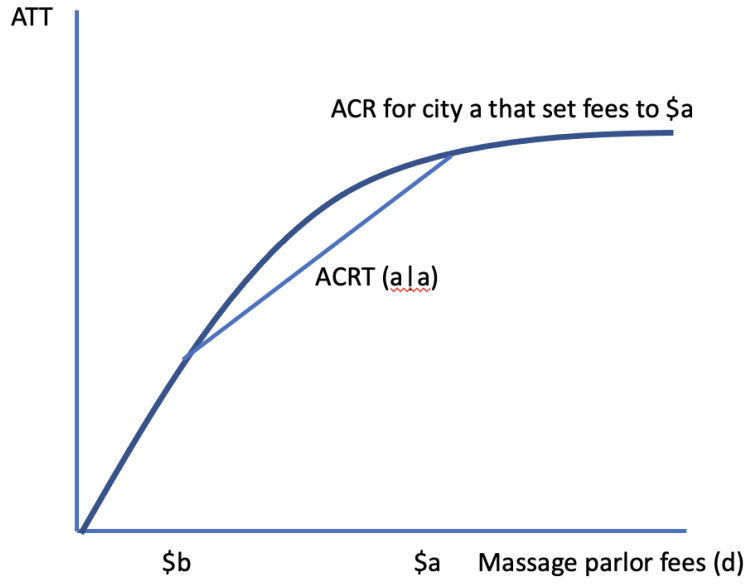
Parameters

Average treated on the treated

$$ATT(d|d) = E[Y_{it}^{d} - Y_{it}^{0}|D_{it} = d]$$

Notice that you are comparing any dose d to no treatment at all – sort of an extensive margin causal response, but that isn't the only causal concept we have. Elasticities are causal, demand curves are causal, but they aren't based on comparisons to nothing – they are intensive margin comparisons, local comparisons, adjacencies. Zero isn't the only counterfactual in other words.

Average causal response for discrete case vs continuous



What is the ACRT?

- ACRT is the causal effect of dose $D = d_j$ vs a different dose $D = D_{j-1}$ for group d
  - → Easiest example is the demand function: at $p = \$10$, I buy 10 units, but at $p = \$11$, I buy 5 units.
  - → Causal effect of that one dollar increase is -5 units
  - → Demand curves are pairs of potential outcomes and treatments and equilibrium "selects" one of them
- Discrete/multi-valued treatment is linear difference between two ATTs for the same city
- Continuous treatment is the derivative of the function itself

Identification for two period set up

1. Random sampling.
2. No anticipation
3. Parallel trends in Y^0 for units of all doses

Identifying $ATT(d|d)$

We can estimate the $ATT(d|d)$ using the simple DiD equation:

$$E[\Delta Y_{it}|D_i = d] - E[\Delta Y_{it}|D_i = 0]$$

No anticipation and parallel trends converts this comparison of before and after into the $ATT(d|d)$

$ATT(d|d)$ is using as its counterfactual the “no treatment”, note. Treatment is a dosage compared to zero iow.

Identifying ACRT

$$\begin{aligned} ATT(b|b) - ATT(a|a) &= (E[\Delta Y_{it}|D_i = a] - E[\Delta Y_{it}|D_i = 0]) \\ &\quad -(E[\Delta Y_{it}|D_i = b] - E[\Delta Y_{it}|D_i = 0]) \\ &= E[\Delta Y_{it}|D_i = a] - E[\Delta Y_{it}|D_i = b] \end{aligned}$$

Comparing high and low dose groups.

Identifying ACRT

$$
\begin{aligned}
ATT(d_j|d_j) - ATT(d_{j-1}|d_{j-1}) &= \\
(ATT(d_j|d_j) - ATT(d_{j-1}|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &= \\
(ACRT(d_j|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &=
\end{aligned}
$$

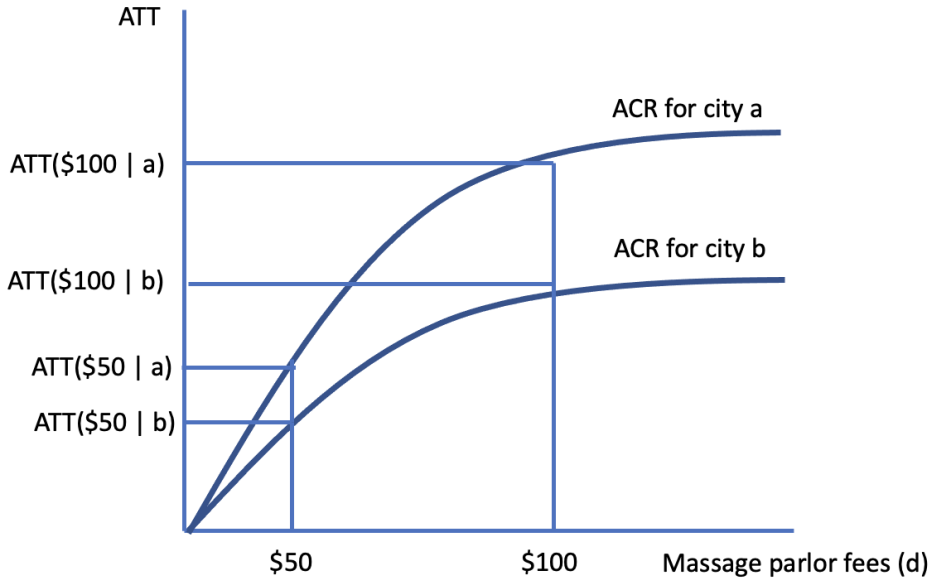
Part in blue is the movement along the average causal response function, the ACRT, and is causal. The part in red is selection bias.

Identifying ACRT

$$
\begin{aligned}
ATT(d_j|d_j) - ATT(d_{j-1}|d_{j-1}) &= \\
(ATT(d_j|d_j) - ATT(d_{j-1}|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &= \\
(ACRT(d_j|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &=
\end{aligned}
$$

Notice parallel trends allows to identify ATT terms but we need additional assumptions for this red part to vanish. We must assume that the ATT for cities that chose d_j and cities that chose d_{j-1} are the same had they both chose d_{j-1}.

Causality and selection bias



Draw the ACRT for top curve and selection bias.

Interpreting this

- Unrestricted heterogenous treatment effects (across dosage levels and across units with difference dose response functions) is not itself the problem
- If we randomized dosages, then $ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1}) = 0$
- Why? Because then there is no selection on gains from dosages, and average causal response functions are the same for all dosage groups
- So then when is this a problem? Sorting on gains

Interpreting this

- When estimating treatment effects using continuous DiD, you will need to make one of two assumptions
  1. Strong parallel trends: Average change in $E[Y^0]$ for the entire sample is the same as the d group
  2. Parallel trends plus homogenous treatment effect functions
- Roy model like sorting on gains typically lead to violations of the second condition insofar as there is heterogenous returns to dosages across units
- So the question you have to ask yourself is do you think that cities are “optimally setting the minimum wage” around some given minimum wage?

Stronger assumption

- I'm really not so sure I think that when it comes to state legislation that I think a Roy model is likely responsible for the equilibrium
- Solving constrained optimization problems is hard and unlikely is it the case that Florida's ATT and Georgia's ATT are terribly different from one another had both chosen the same minimum wage (but that is the bias)
- But notice where this is taking us – we are getting closer and closer back to a place of assuming away the problems!

Interpretation of estimate

- So we are back to interpretation – what is the interpretation, then, of $\widehat{\delta}^{TWFE}$ with continuous DiD?
- Callaway, Goodman-Bacon and Sant'anna (2022) have a decomposition which I encourage you to read

TWFE as weighted average of ACRT parameters

$$\widehat{\delta}^{TWFE} = \int_{d=L}^{d=U} w_1(l) ACRT(l), dl + w_0 \frac{ATE(d_L)}{d_L}$$

- They assume strong parallel trends and find that TWFE coefficient is weighted average of ACRT function
- All weights are non-negative and sum to one
- But weights are strange in that they are maximized at dosage equal to the mean dosage of the entire dataset

Conclusion

- Very interesting area, a bit challenging, my suggestion is make clear your assumptions
- Most intuitive for me is the parallel trends plus homogenous treatment effects across units around the dosages
- But interpretations are hard because you're having to think about a new parameter, and you could have nonlinearities in the ACRT that since TWFE is a weighted average of them, could flip sign
- Not a problem but could be a challenge for you as you try to make sense of it (particularly given how the weights are)
- Unclear to me how we might develop original falsification exercises for this homogeneity assumption except it does imply equalities (covariate balance?)

Roadmap

Continuous DiD

- Dose causal parameter
- Identification
- Selection bias
- Interpreting TWFE

Concluding remarks

Empirical micro model

- In 2014, David Card gave a speech at Michigan in which he told the history of the “empirical micro model”, particularly labor
- He notes the different role that economic models have played throughout the history of modern applied (empirical) work
    1. Simple approximations of the theoretical model (regress earnings on education)
    2. Structural modeling (often highly parametrized models)
    3. Princeton (essentially this entire workshop) where focus is on physical (manipulated) treatment assignment
- Models are valuable even under #3, but largely to suggest question, interpret results – not to solve the identification problem

Unrestricted heterogeneity and selection on gains

- Over time, applied economists (at least the modal one) is far less likely to believe they know enough about treatment assignment or the underlying treatment effects that they are willing to make any theoretical restrictions
- Treatment effects come from unknown production functions, and whether we have heterogeneity across treatment groups involves selection, most likely from a Roy model type of situation
- By and large, this is my conjecture, economists are either unwilling to restrict heterogeneity ex ante, and the only models many of us are by and large committed to are Roy like rational sorting models
- Without restrictions on heterogeneity ex ante, Roy models imply sorting on gains which create the very problems we find

Consequences of being “deeply a-theoretical”

- Insofar as people sort on the gains from treatment, then given unknown and unrestricted heterogeneous treatment effects, we do not know enough to rule out situations like those in our “baker” dataset
- Baker dataset was extreme, but before this literature, it’s not clear anyone would’ve known it was extreme
- Linear dynamics that never stabilized, large treatment effects in the ATT, heterogenous treatment profiles
- TWFE in the canonical specification performs poorly and in our case flipped sign

Decision making under uncertainty

- If you cannot restrict heterogeneity, and Roy like models are driving selection into treatment, then there's risks
- TWFE will be efficient if model is correctly specified, which requires no dynamic treatment effects, and most of us when pushed cannot credibly claim we have enough prior knowledge to say that as we are often working on something for which underly production technology around treatment effects are not well known or external validity is unknown
- You will need to take this seriously, and adjust practices *away* from naive TWFE modeling
- You are, though, the first wave of researchers to have to, so you are unfortunately caught in a bit of limbo moment

Future work

- We are going to cover more differential timing modeling, but I think some of this is a bit redundant
- We start to see similar fixes to the underlying problem performed in different, though internally valid, ways
- Much of this is a bit cloudy to even the most up-to-date practitioner – differences between estimators are based on which control groups are used, whether PT holds for those control groups, and how best to think about the parallel trends assumption (some of which have stronger additive forms imposing pre-treatment parallel trends and others which do not)
- Cherry picking diff-in-diff could be on the horizon, and this workshop is meant to prepare you for this situation as worst case scenario you will be reading referee reports of people using methods which you may yourself think are inferior but which you still need to understand
- You will need to continue to invest in good judgment, as always; most likely going forward returns to econometric skill has shifted up

Suggestions

- These papers have made the careers of young econometricians to be honest – the growth in citations in short period of time is very weird for econometrics papers (even the IV papers by Angrist and Imbens don't show this pattern)
- Think of a monopolistic competition model – entry stops when zero profits at the margin
- **Profits are shrinking but not zero at the margin**
- There's a lot of activity, it will continue, and you'll need to allocate your time accordingly
- Good luck with your research and pursuit of meaning and happiness – don't lose sight of what's important