Introduction

Staggered Rollout

Bacon Decomp

Resumiendo

Los nuevos Estimadores

Relaxing Parallel Trends

Inferencia

Otros topicos

DiD: foundations and new methodology. Dia 2

Sebastian Tello-Trillo[1]

[1]University of Virginia & NBER, UDEP

Lima Summer School in Economics 2023

Difference-in-Difference: The new stuff

- La nueva literatura de DiD se ha esta moviendo rapidamente, y a veces es confuso manejar el panorama.
- Recomiendo leer "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature" de Jonathan Roth, Pedro H. C. Sant'Anna, Alyssa Bilinski, John Poe
- En si, la literatura reciente se enfoca en extender los supuestos de DiD para otros casos que son mas parecidos a los que uno encuentra en policy evaluation
- En si, esto fue lo que se convertiera en una "Revolucion", es porque la practica se habia extendido mas rapido que la teoria.

DiD: The new stuff

- ¿Como podemos entender la nueva literatura?
- La organizaremos en 4 categorias
  - Multiple periods and variation in treatment timing
  - Investigating parallel trends (conditional on covariates, can we still learn something?)
  - Alternative Sampling Assumptions (Inference)
  - Other topics (Continous treatment, Spillovers, heterogenous effects).

Multiple periods and variation in treatment timing

- Cambios en la Notacion:
- $D_{i,t}$ indica que la unidad i recibe el tratamiento en el periodo t.
- El periodo mas temprano que la unidad i recibe el tratamiento es$G_i = min\{t : D_{i,t} = 1\}$
- Por ende si una unidad no recibio tratamiento, que seria G?
  - $G_i = \infty$
- Treatment es un "absorbing state". Osea solo se "turn on" y no se turn off.
  - $D_{i,t} = 1 \; \forall \; t \geq G_i$

Notacion

- T=en numero total de periodos
- Antes teniamos $Y_{i,t}(0,0)$, ahora como puedes haber mas periodos:
  - $Y_{i,t}(g) = Y_{i,t}(0_{g-1}, 1_{T-g+1})$. ¿Que significa eso?
  - Que significa $Y_{i,2016}(2014)$
  - Como seria una unidad que nunca fue tratatada?
    - $Y_{i,t}(\infty) = Y_{i,t}(0_T)$
- Notese que en el mundo de 2x2, la notacion hubiera sido $Y_{i,t}(2)$ enves de $Y_{i,t}(0,1)$ y $Y_{i,t}(\infty)$ enves de $Y_{i,t}(0,0)$.

Extendiendo DiD

- Si tenemos mas periodos, y solo un grupo de tratamiento que cambia en un periodo especifico, en si los supuestos son los mismos que antes.
- Si tenemos mas periodos, y un grupo (>1) de tratamiento que cambia en el mismo periodo, los supuestos son los mismos que antes
- Si tenemos mas periodos, y un grupo de tratamiento que cambia en distintos tiempos, las cosas cambian un poco.

Staggered Adoption: Assumptions

- ¿Como cambian los supuestos?
- No anticipation, sigue siendo lo mismo
  - $Y_{i,t}(g) = Y_{i,t}(\infty)\ \forall_i\ y\ t > g$
- Supuesto de PT para staggered settings: For all $t \neq t'$and $g \neq g'$

$$\mathbb{E}[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g] = \mathbb{E}[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g']$$

- Este supuesto impone que en el contrafactual donde no se hubiera producido el tratamiento, los resultados promedio para todos los grupos que adoptaron la reforma habrían evolucionado en paralelo.
- Por ejemplo, si no hubiera habido reformas, los outcomes hubieran evolucionado en paralelo (en promedio) a los grupos que tuvieron la reforma y a los que nunca tuvieron la reforma
- Para entender esto bien, vamonos a un grafico.

Staggered Adoption: Assumptions

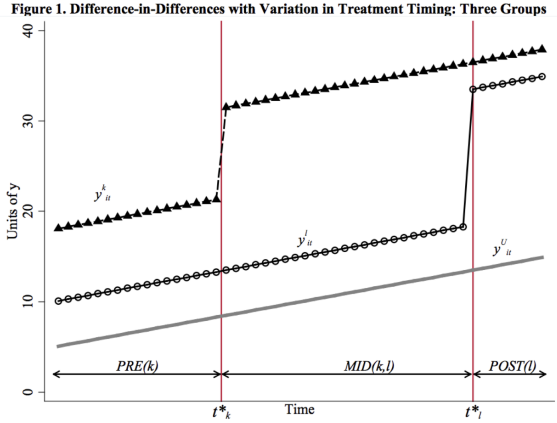
- ¿Como cambian los supuestos?
- Supuesto de PT para staggered settings: For all $t \neq t'$and $g \neq g'$

$$\mathbb{E}[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g] = \mathbb{E}[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g']$$

- Este supuesto impone que en el contrafactual donde no se hubiera producido el tratamiento, los resultados promedio para todos los grupos que adoptaron la reforma habrían evolucionado en paralelo.
- Por ejemplo, si no hubiera habido reformas, los outcomes hubieran evolucionado en paralelo (en promedio) a los grupos que tuvieron la reforma y a los que nunca tuvieron la reforma
- Para entender esto bien, vamonos a un grafico.

Staggered Adoption

- Que significa que estan adoptando en diferentes tiempos?

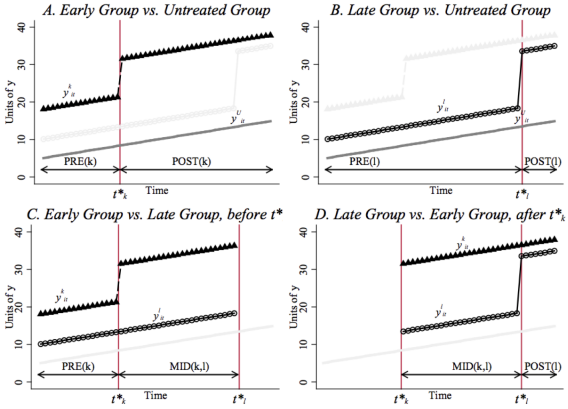


Bacon's Decomposition

- *Difference-in-Differences with Variation in Treatment Timing (2021) Andrew Goodman-Bacon. Journal of Econometrics*
- Si usamos un TWFE modelo para estimar esto, que nos da el beta asociado al parametro DiD?

Bacon's Decomposition

Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



- Resulta que TWFE DD estimator (pooled) es el promedio ponderado (weighted average) de todas las comparacion 2x2.

$$\widehat{\beta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\beta}^{2x2}_{kU} + \sum_{k \neq U} \sum_{\ell > k} \left[s^{k}_{k\ell} \widehat{\beta}^{2x2,k}_{k\ell} + s^{\ell}_{k\ell} \widehat{\beta}^{2x2,\ell}_{k\ell} \right]. \qquad (10a)$$

Where the 2x2 DD estimators are:

$$\widehat{\beta}^{2x2}_{kU} \equiv \left(\overline{y}^{POST(k)}_{k} - \overline{y}^{PRE(k)}_{k} \right) - \left(\overline{y}^{POST(k)}_{U} - \overline{y}^{PRE(k)}_{U} \right), \qquad (10b)$$

$$\widehat{\beta}^{2x2,k}_{k\ell} \equiv \left(\overline{y}^{MID(k,\ell)}_{k} - \overline{y}^{PRE(k)}_{k} \right) - \left(\overline{y}^{MID(k,\ell)}_{\ell} - \overline{y}^{PRE(k)}_{\ell} \right), \qquad (10c)$$

$$\widehat{\beta}^{2x2,\ell}_{k\ell} \equiv \left(\overline{y}^{POST(\ell)}_{\ell} - \overline{y}^{MID(k,\ell)}_{\ell} \right) - \left(\overline{y}^{POST(\ell)}_{k} - \overline{y}^{MID(k,\ell)}_{k} \right). \qquad (10d)$$

Bacon's Decomposition

The weights are:

$$s_{kU} = \frac{(n_k + n_U)^2 \overbrace{n_{kU}(1 - n_{kU})\overline{D}_k(1 - \overline{D}_k)}^{\hat{V}^D_{kU}}}{\hat{V}^D}, \tag{10e}$$

$$s^k_{k\ell} = \frac{\left((n_k + n_\ell)(1 - \overline{D}_\ell)\right)^2 \overbrace{n_{k\ell}(1 - n_{k\ell})\frac{\overline{D}_k - \overline{D}_\ell}{1 - \overline{D}_\ell}\frac{1 - \overline{D}_k}{1 - \overline{D}_\ell}}^{\hat{V}^{D,k}_{k\ell}}}{\hat{V}^D}, \tag{10f}$$

$$s^\ell_{k\ell} = \frac{\left((n_k + n_\ell)\overline{D}_k\right)^2 \overbrace{n_{k\ell}(1 - n_{k\ell})\frac{\overline{D}_\ell}{\overline{D}_k}\frac{\overline{D}_k - \overline{D}_\ell}{\overline{D}_k}}^{\hat{V}^{D,\ell}_{k\ell}}}{\hat{V}^D}. \tag{10g}$$

and $\sum_{k \neq U} s_{kU} + \sum_{k \neq U} \sum_{\ell > k} \left[s^k_{k\ell} + s^\ell_{k\ell}\right] = 1.$

Bacon's Decomposition

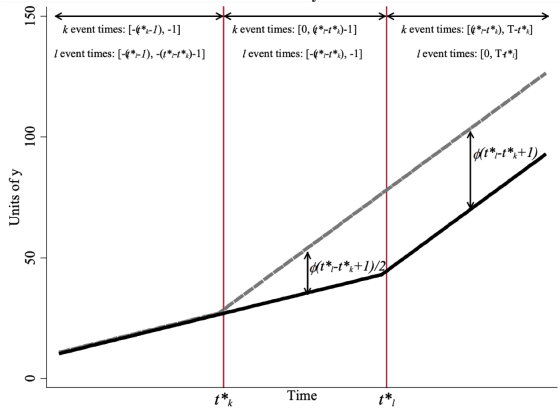
- De donde vienen los pesos? Los pesos provienen de tres elementos:
  - Tamaño de cada subgrupo en base a N de los grupos: qué porcentaje de unidades - estados, condados, empresas, trabajadores, etc.- se encuentran en el grupo de tratamiento y control para un par (2x2) dado.
  - Tamaño de cada subgrupo en base al tiempo: qué porcentaje de períodos de tiempo se utilizan en una submuestra 2x2 determinada).
  - La varianza del tratamiento (qué tan cerca del principio/final de la ventana de submuestra se activa el tratamiento.
- Implica que: Estos pesos terminan otorgando un alto grado de peso a las unidades tratadas en el medio de la muestra (Ya que tienen la varianza más alta en el indicador de tratamiento)

Bacon's Decomposition

- ¿Y para que me sirve saber que el TWFE hace eso con staggered timing? ¿Que gano con esto?
- Uno para entender mejor que es lo que hace OLS
- Dos para entender a que le esta poniendo peso mas que a otras cosas
- Tres, si crees que los efectos de tu tratamiento son instantáneos e invariantes en el tiempo (time invariant),los pesos son todos positivos. Todo fresh.
- Cuatro: Pero si los efect varian a traves del tiempo, esto puede llevar a pesos negativos, y crea un cesgo en el estimado del DiD.

Bacon's Decomposition

Figure 3. Difference-in-Differences Estimates with Variation in Timing Are Biased When Treatment Effects Vary Over Time



Bacon's Decompositions

- Goodman-Bacon provides a way to assess the weights in a given TWFE design.
- Worksheet #3
- Negative weights tend to occur for early-treated units in periods late in the sample ("Forbiden Comparisons")
- Entonces una implicacion importante de este trabajo no es que "TWFE" este "wrong" sino que cuando los efectos son variantes a traves del tiempo que uno mide (short is better?) en staggered treatment puede que haya un problema de usar TWFE.

Another way

- Hay otra manera mas sencila de ver esto. Si entendemos como se obtiene un beta a traves de Frisch-Waugh-Lovell theorem (worksheet #4).

$$\widehat{\beta}_{post} = \frac{\sum_{i,t}(D_{i,t} - \widehat{D}_{i,t})Y_{i,t}}{\sum_{i,t}(D_{i,t} - \widehat{D}_{i,t})^2}.$$

Regresando al supuesto

- Supuesto de PT para staggered settings: For all $t \neq t'$and $g \neq g'$
$$\mathbb{E}[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g] = \mathbb{E}[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g']$$
- Este supuesto impone que en el contrafactual donde no se hubiera producido el tratamiento, los resultados promedio para todos los grupos que adoptaron la reforma habrían evolucionado en paralelo.
- Por ejemplo, si no hubiera habido reformas, los outcomes hubieran evolucionado en paralelo (en promedio) a los grupos que tuvieron la reforma y a los que nunca tuvieron la reforma

Nuevos Supuestos

- En los nuevos papers, hay versiones diferentes (menos estricta) que esta version del supuesto.
- Por ejemplo: Difference-in-Differences with Multiple Time Periods (Callaway & Sant'anna, 2021), relaja el supuesto, y solo impone que PT para los años despues de la reforma para alguna de las unidades tratadas.

Assumption 4.a (Parallel trends for staggered setting - post-treatment only)**.**

$$\mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t'}(\infty) \,|\, G_i = g\right] = \mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t'}(\infty) \,|\, G_i = g'\right].$$

for all $t, t' \geqslant g_{min} - 1$*, where* $g_{min} = \min \mathcal{G}$ *is the first period where a unit is treated.*

Nuevos Supuestos

- Esto require que los grupos que recibieron una reforma en distintos tiempos tengan PT solo despues de la reforma, pero no necesariamente antes de la reforma.
- Sun and Abraham (2021) tiene un caso donde solo impone esto a los grupos que serian tratatos (ever treated) y no impone esto a los grupos nunca-tratados.
  - Esto quiere decir PT entre todos los grupos que adoptarian la reforma, pero no entre los grupos que adoptarian y los que nunca adoptaron.

Resumiendo

- En si entonces, teniendo staggered timing, los supuestos cambian ever-so slightly, y puedes ser relajados o no.
- Pero son analogos al DiD 2x2
- ¿Entonces cual es el problema? ¿Por que la literature se ha hecho un problema?
- No es por los supuestos, pero por la estimacion. Como vemos en el ejercicio de Bacon, la estimacion cuando haya efectos hetereogeneos a traves del tiempo es el problema.
- Entonces, si no podemos usar TWFE para estimar, que alternativas, hay?

Problems with TWFE

- Los sgte papers empezaron a mostrar de diferentes maneras los problemas de TWFE
- Revisiting Event Study Designs: Robust and Efficient Estimation (Kirill Borusyak, Xavier Jaravel, Jann Spiess) RR ReSTUD
- Two-way Fixed Effects and Differences-in-Differences Estimators with Several Treatments (Clément de Chaisemartin and Xavier D'Haultfoeuille) *Journal of Econometrics* (2022/2023)
- *Difference-in-Differences with Variation in Treatment Timing (2021) Andrew Goodman-Bacon. Journal of Econometrics*

Problems with Event Studies

- Sun y Abraham (2021) investiga si el problema de TWFE ocurre tambien en los event studies.
- Ellos muestran que cuando hay efectos de tratamiento dinámicos heterogéneos en las cohortes de adopción, los coeficientes del ES se vuelven difíciles de interpretar.
- Y aunque puedan que hallan PT, los ES no te mostraran eso (caso que vimos en el worksheet #4).
- En si los problemas del ES son parecidos, y vienen del problema de los negative weights.

Soluciones a estos problems

- Diagnosticos:
  - Goodman-Bacon (2021) propone informar los pesos que el parametro DiD (2x2), lo que permite evaluar cuánto peso se le da a las “comparaciones prohibidas" de unidades ya tratadas y cómo eliminar las comparaciones cambiaría la estimación.
  - de Chaisemartin y D'Haultfoeuille (2020) proponen informar el número/fracción de los cortes ATT que reciben ponderaciones negativas, así como el grado de heterogeneidad en los efectos del tratamiento que sería necesario para que el efecto del tratamiento estimado tenga el "signo incorrecto".
  - Pamela Jakiela (2021) propone evaluar tanto si TWFE otorga ponderaciones negativas a algunas unidades tratadas como si los datos rechazan el supuesto de efectos de tratamiento constantes.

Nuevos estimadores

- The Callaway and Sant'Anna estimator. (CS, 2021)
- Imputation estimators. Borusyak, Jaravel and Spiess (BJS, 2021)
- Otros

Callaway and Sant'Anna estimator (2021)

- Comienzan con la idea que los ATT varian a traves de el tiempo y el grupo (cohorte).

$$ATT(g,t) = E[Y_{i,t}(g) - Y_{i,t}(\infty)|G_i = g]$$

- Que da el average treatment effect on the treated, para el tiempo t, para el cohorte que fue tratado por primera vez en el tiempo g.
- ¿Que es ATT(2014, 2016)?
- Con los supuestos generalizados (relajados) del staggered DiD, ellos pueden identificar el $ATT(g,t)$ comparando el cambio que paso, con el cambio que hubiera pasado al corte g, entre g-1 y t.
- El cambio que hubiera pasado, se obtiene de un grupo que no ha sido tratado hasta el periodo t.
- En si solo comparaciones "limpias".

Callaway and Sant'Anna estimator (2021)

- $ATT(g,t) = E[Y_{i,t} - Y_{i,g-1}|G_i = g] - E[Y_{i,t} - Y_{i,g-1}|G_i = g']$ *for any* $g' > t$
- $ATT(2014, 2016) = E[Y_{i,2016} - Y_{i,2013}|G_i = 2014] - E[Y_{i,2016} - Y_{i,2013}|G_i = 2017+]$ *for any* $g' > t$
- De ahi solo se reemplazan los expectations por el sample analog (promedios!)
- En si el estimador puedes escoger que solo use los (1) never treated o los (2) not yet treated.
- Tambien puedes hacer que haga un promedio de los dos. (Cuando los periodos son pocos).

Callaway and Sant'Anna estimator (2021)

- Pero cuando hay un monton de grupos y tratamientos tendriamos un monton de $ATT(g,t)$!
- CS proponen resumir esos parametros haciendo un promedio ponderado de varios de esos $ATT(g,t)$.
- Los pesos podrían elegirse para ponderar diferentes cohortes por igual, o en términos de sus frecuencias relativas en la población tratada.
- En el paper discuten un numero de "opciones" de los pesos y las interpretaciones que se les puede dar. Esto es parte de la eleccion del investigador.
- Inferencia se hace a traves de bootsrapping.

Introduction

Staggered Rollout

Bacon Decomp

Resumiendo

Los nuevos Estimadores

Relaxing Parallel Trends

Inferencia

Otros topicos

Aplicacion

- R & STATA csdid (csdid2).

Inputation Estimator: BJS

- Borusyak, Jaravel and Spiess (2021) introduce a related approach which they refer to as an imputation estimator (see, also, Gardner (2021), Liu, Wang and Xu (2022) and Wooldridge (2021) for similar proposals.
- ¿Que es lo que proponen?
- Paso 1: Correr una regresion $Y_{i,t}(\infty) = \alpha_i + \lambda_t + \varepsilon_{i,t}$
  - ¿Que significa esto?
  - Correr una regresion solo para los never-treated contra los efectos fijos de tiempo y grupo.
  - Usando los valores de esta regresion $\hat{\alpha}_i + \hat{\lambda}_t$ para crear $\hat{Y_{i,t}}(\infty)$.

Inputation Estimator: BJS

- Estos $\hat{Y_{i,t}}(\infty)$ son los que usaremos como los counterfactuals de los *grupos tratados*, bajo supuestos parecidos a los que vimos antes.
- La estimacion se vuelve algo sencillo: $Y_{i,t}(g) - \hat{Y_{i,t}}(\infty)$
- Estas diferencias "individuales" (de grupo), se agregan en parametros que se resumen en versiones de $ATT(g,t)$

La diferencia entre BJS y CS?

- Para entender la diferencia entre estos metodos, usemos un ejemplo sencillo
- Imaginemosno un caso en el que hayan 3 periodos $t = 1, 2, 3$
- Algunos grupos seran tratados en el periodo 3
- Otros nunca son tratados $t = \infty$
- Eso quiere decir que $G = \{3, \infty\}$

La diferencia entre BJS y CS?

- En este setting simple, cuales son las diferencias.
- CS estimara $ATT(3,3)$, y esto se obtiene a traves
  - $\widehat{ATT}_{cs}(3,3) = (\bar{Y}_{3,3} - \bar{Y}_{3,\infty}) - (\bar{Y}_{2,3} - \bar{Y}_{2,\infty})$
- Donde $\bar{Y}_{t,g}$ es el promedio de Y, en el periodo t, para las unidades with $G_i = g$
- $\bar{Y}_{2,3}$, es el promedio de Y en el periodo 2 para las unidades que fueron tratadas en $t = 3$
- $\bar{Y}_{2,\infty}$, es el promedio de Y en el periodo 2 para las unidades que nunca fueron tratadas, $t = \infty$

La diferencia entre BJS y CS?

- En este setting BJS va a user la "imputacion" del counterfactual.
  - $\widehat{ATT}_{bjs}(3,3) = (\bar{Y}_{3,3} - \bar{Y}_{3,\infty}) - (\bar{Y}_{pre,3} - \bar{Y}_{pre,\infty})$
- Donde $\bar{Y}_{pre,3}$ es el promedio de Y, en los periodos antes de el tratamiento $(t = 1,2)$, para las unidades que fueron tratadas en $t = 3$
- $\bar{Y}_{pre,\infty}$ es el promedio de Y, en los periodos antes de el tratamiento (t=1,2), para las unidades que no fueron tratadas.

La diferencia entre BJS y CS?

- Entonces la gran diferencias es que:
- CS realiza todas las comparaciones en relación con el último período de pretratamiento
- mientras que BJS realiza comparaciones en relación con el promedio de los períodos de pretratamiento.
- Y ¿Cual es mejor?

La diferencia entre BJS y CS?

- Para decidir cual es mejor, esto depende de los "tradeoffs", en si los tradeoffs son (1) eficiencia del estimador y (2) los supuestos.
- BJS va a ser mas eficiente, porque esta agregando mas informacion (promediando a traves de varios pre-treatment period). Con solo "no anticipation". BJS is efficiente con homoskedasticity y serially correlated errors.
    - Estos no pueden que sean los setting normales, pero en la practica esto significa que en settings donde no haya mucha heteroskedastic o serially correlated erors BJS va a ser mas eficiente que CS.
- CS usa un supuesto mas "relajado" o "flexible" que BJS. CS solo requiere PT entre el periodo 2 y 3, solo PT entre el antes y el despues (una ventana mas corta).
- BJS requiere PT para el average outcome de 1,2 y 3.
- Pensemos en esto cuando el setting se vuelve mas complicado.

La diferencia entre BJS y CS?

- Entonces generalizando el problema, BJS requiere PT entre todos los grupos y periodos (antes y despues) y CS requiere PT solo para los periodos "post-treatment".
- La pregunta entonces es:
  - PT, son validas por cuanto tiempo en mi caso?
  - Por ejemplo, si el Y potencial "sin tratar" hubiera aumenta más rápido entre las unidades tratadas que entre las unidades no tratadas a lo largo del tiempo, entonces la violación de PT es mayor cuando comparamos períodos más separados. Esto crearia mas un cesgo en BJS que en CS.

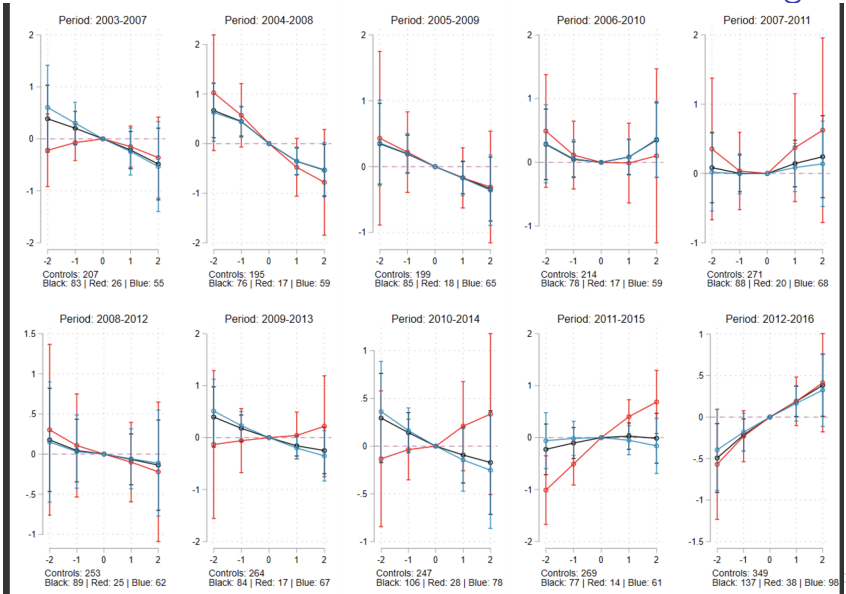
Otros estimadores

- de Chaisemartin and D'Haultfoeuille (2020) y CS en el setting de staggered timing con absorbing state, sonm equivalentes.
  - Pero se puede usar cuando el tratamiento es on/off
- Sun and Abraham (2021), parecido a CS, pero el mismo metodo escoge o never-treated o not-yet treated, con preferencia a los never-treated.
- Cengiz, Dube, Lindner and Zipperer (2019): Stacked Regression

Stacked Regression

- La idea del stacked-regression (formalizado por Gardner (2021)). Es crear diferentes muestras de manera manual, y hacer "simple" DiD en cada una de esas comparaciones a traves de TWFE
- Despues se podrian "agregar" los beta, pero el problema es que no hay pesos.
- Para poder crear pesos, se corre la regresion en conjunto y se crean "cohort-by-group" y "cohort-by-time" Fixed effects.

Stacked Regression



Introduction

Staggered Rollout

Bacon Decomp

Resumiendo

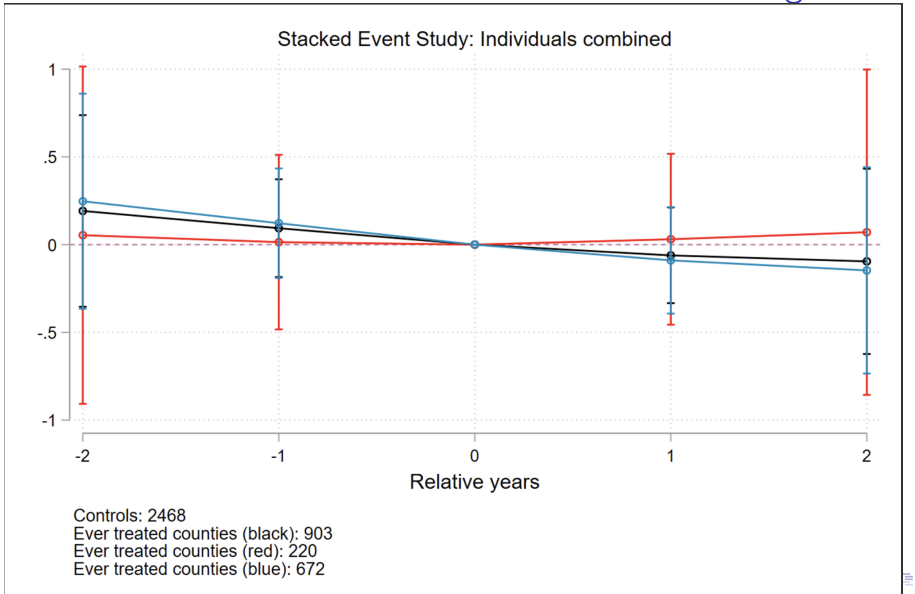
Los nuevos Estimadores

Relaxing Parallel Trends

Inferencia

Otros topicos

Stacked Regression



Otras Extensiones

- ¿Que pasa cuando el tratamiento se prende y se apaga?
  - de Chaisemartin and D'Haultfoeuille (2020) and Imai and Kim (2021) consideran estos casos.
  - Estos estimadores comparan los cambios en Y de las unidades cuyo estado de tratamiento cambió (off) con otras unidades cuyo estado de tratamiento se mantuvo constante durante los mismos períodos.
  - Pero para interpretar este parametro se necesitan algunos supuestos como el "no carryover" supuesto.
  - Este impone que los Y "que hubieran ocurriso" dependan solo del estado actual del tratamiento y no del historial completo del tratamiento.

Otras extensions

- Ejemplo: si el tratamiento es un aumento en el salario mínimo y el Y es el empleo, entonces el supuesto de "no carryover" requiere que el empleo en el período t dependa solo de si el salario mínimo se aumentó en el período t y no del historial de cambios en el salario mínimo.
- de Chaisemartin and D'Haultfoeuille (2022) esta tratando de relajar el supuesto (y veremos si funciona con el estimador).
  - El supuesto de este paper permiten que los Y "que hubieran ocurrido" dependan de la ruta completa de los tratamientos y, a cambio, imponen que un PT más fuerte. (Que sean PT en todo el periodo).

Otras extensions

- Non-binary treatment o "dosis" o continous DiD
- de Chaisemartin and D'Haultfoeuille (2018) study "fuzzy" DiD
- Callaway, Goodman-Bacon and Sant'Anna (2021): continuous DD
  - Solo supuestos e interpretatciones, ningun nuevo estimador.
  - Supuestos: que el efecto sea lineal

Recomendaciones en la practica

- Saber que tipo de comparaciones esta haciendo tu estimador
- Es mejor construir los resultados de los comandos por uno mismo y despues usar el comando
  - Esto ayuda entender lo que uno esta haciendo.
- Los event study (even with TWFE) te pueden decir bastante de las comparaciones. Es dificil crear situaciones en las que los ES se vean "bien" y haya heterogenidad del efecto a traves del tiempo.
  - Si las pre-trends no se ven bien, use el bacon decomposition para entender si es el problema es la heterogenidad.
  - Si es eso, utiliza un nuevo estimador, sino puede que sea una mala comparacion.
- O tb puedes usar los ES con TWFE, solo con los never-treated. O solo con los "will-be-treated".

Recomendaciones practicas

- ¿Cual estimador uso?
  - En la practica uno podria usar varios (worksheet #8)
  - Si todos te dan la misma respuesta podrias esocger uno
  - Si te dan respuestas diferentes, tendriamos que entender ¿por que?
  - Como el caso de BJS vs. CS, entender como las pequeñas diferencias de los estimadores estan resultando en cosas diferentes, esto quiere decir que habra un contexto en el que uno de ellos sea mas adecuado.

Relaxing Parallel Trends Assumptions

- Ahora vamos a ir al segundo gran punto de la literatura: relajando los supuestos de PT.
- Recordemos que:
  - A veces aunque haya evidencia empirica, no vamos a estar muy seguros si el supuesto de PT sea cierto o no.
  - Los PT son sensibles a la funcion del Y (Functional form). Si es lineal, no se sabe si funcionaria con log o poisson.

Relaxing PT Assumptions

- Una forma de "relajar" el supuesto, es que creemos en el supuesto una vez que hayamos condicionado por una variables de control.
- Esto nos ayudaria bastante a llegar a "creer" PT, porque ahora el supuesto no es "crecerian de las misma manera" sino "conditional al ingreso de la muni, el partido politico, y otras cosas, esta y otra muni hubieran cambiado de la misma manera".
- Queda claro que solo queremos poner "pre-treatment" controls, y no controls que tambien tengan valores post-treatment.

PT conditional on Covariates

- El supuesto de PT entonces cambiaria a:

$$E[Y_{i,2}(0,0) - Y_{i,1}(0,0)|D_i = 1, X_i] = E[Y_{i,2}(0,0) - Y_{i,1}(0,0)|D_i = 0, X_i]$$

- En el staggered timing, el supuesto seria analogo.
- Este supuesto, tambien require del "overlap condition" o "positivity condition"
  - Que garantiza que por cada unidad tratada con co-variables X_i, hay al menos algunas unidades no tratadas en la población con el mismo valor de X_i
  - Similar a common support de matching. Esto se verifica empiricamente.

PT conditional on Covariates

- Con las condiciones de PT, no anticipation, y overlap, entonces el efecto causal seria:

$$\tau_2(x) = E[Y_{i,2} - Y_{i,1}|D_i = 1, X_i = x] - E[Y_{i,2} - Y_{i,1}|D_i = 0, X_i = x]$$

- Usando la law of interated expectations

$$\tau_2 = E[Y_{i,2}(0,1) - Y_{i,2}(0,0)|D_i = 1 = E\left[E[Y_{i,2}(0,1 - Y_{i,2}(0)|D_i = 1|X_i]|D_{i=1}\right]$$

- En sencillo, esto quiere decir que podemos recuperar *el unconditional ATT* al promediar el ATT condicional con toda la distribution de X_i en la poblacion.

Estimacion: PT Conditional on Covariates

- En el caso de TWFE, esto require interactuar las X's con los efectos fijos de grupo y de tiempo (por separado).
- Otros ajusted parecidos son: Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998).
- Match unidades tratadas con unidades no tratadas de covariantes similares, y correr TWFE.
- IPW: Usar las covariates para construir la probabilidad de que una unidad sea tratada.

Estimacion: PT Conditional on Covariates

- $\tau_2 = \frac{E\left[\left(D_i - \frac{(1-D_i)p(X_i)}{1-p(X_i)}(Y_{i,2} - Y_{i,1})\right)\right]}{E[D_i]}$
- Sant'Anna and Zhao (2020). Double Robust Estimators:
  - Utiliza el IPW + Regression
  - csdid utiliza drdid.
- En si la mayoria de los comandos que hemos visto, tienen la flexibilidad de incluir covariantes, y es simplemente cuestion de ponerlo en las opciones.

Time-varying covariates?

- Caetano, Callaway, Payne and Rodrigues (2022) discuten este caso. En si lo que uno tiene que asumir es que las covariates no estan siendo afectadas por el tratamiento, pero si afectan la seleccion del tratamiento.
- Lo principal es que no sean outcomes.

Pre-reform outcomes?

- Ryan (2018) propone que controlar por Y rezagados (lagged) puede reducir el sesgo de los unobserved confounding.
- Daw and Hatfield (2018) argumenta que tienen que haber condiciones especificas sino se puede cesgar el estimador.
  - Cuando los grupos de tratamiento y de control tienen diferentes distribuciones de Y, pero las mismas tendencias, hacer matching de los grupos de tratamiento y control en los Y rezagados, selecciona unidades de control con un "shock" particularmente grande en el período previo al tratamiento (en si agarra Ys de una parte "extrema" de la distribucion, para que se parezca al grupo de tratamiento).
  - Esto puede inducir un sesgo debido a un efecto de reversión a la media, cuando de hecho no condicionar los resultados retrasados habría producido tendencias paralelas.

Testing for pre-existing trends

- The Promise and Pitfalls of Differences-in-Differences: Reflections on '16 and Pregnant' and Other Applications Kahn-Lang and Lang (2020):
  - Incluso si las tendencias previas son exactamente paralelas, esto no garantiza necesariamente que se satisfaga la suposición de tendencias paralelas posteriores al tratamiento.
  - Ejemplo: la estatura promedio de niños y niñas evoluciona en paralelo hasta alrededor de los 13 años y luego diverge, pero no debemos concluir de esto que existe un efecto causal de los bar mitzvahs (que ocurren para los niños a los 13 años) en la altura de los niños.

Testing for pre-existing trends

- Bilinski and Hatfield, 2018; Freyaldenhoven, Hansen and Shapiro, 2019; Kahn-Lang and Lang, 2020; Roth, 2022
- Indican que estos "tests" tienden a encontrar coeficientes no significativos pero esto es porque son test de "low power"
- ¿Que es "power"? es la probabilidad de que una prueba de hipótesis pueda detectar un efecto en una muestra cuando el efecto si existe en la población
- Por ejemplo, si el poder estadístico es del 80 %, una prueba de hipótesis tiene un 80% de posibilidades de detectar un efecto que realmente existe. Ahora imagine que está realizando un estudio que tiene solo el 10%. Eso no es bueno porque es mucho más probable que la prueba no detecte el efecto.

¿Por que low power?

- En si porque los betas estan diseñados a testear si algo es 0 (null) o no (fail to reject null), no estan diseñados a testear si esos beta son grandes (i.e si es que hay una diferencia) o no.
- Los test solo "reject the null" cuando hay bastante evidencia en contra de la hipotesis nula ($\beta = 0$). Entonces tiende a no-reject.
- Roth (2022) encuentra que el poder de los tests en papers que ya han sido publicados en top 5 economic journals es the 50%.

Diagnosticos

- El power de un test cambia de acuerdo al contexto y la data.
- Roth (2022) desarrolló herramientas para realizar análisis de poder y calcular las distorsiones probables de la prueba previa bajo la hipótesis de violaciones de tendencias paralelas planteadas por el investigador.
- Estas herramientas permiten al investigador evaluar si es probable que las limitaciones descritas anteriormente sean severas para posibles violaciones de tendencias paralelas consideradas económicamente relevantes.
- Link al test.

Violations of PT but partial identification

- ¿Que pasa si no encontramos evidencia de PT en el pre-test?
- En si lo usual es (1) cambiar el grupo de control, (2) buscar covariantes, y (3) rendirse :(
  - Rambachan y Roth (2022b) proponen un enfoque para la inferencia robusta y el análisis de sensibilidad cuando se pueden violar las tendencias paralelas, basándose en trabajos anteriores de Manski y Pepper (2018).
  - La idea basica es sino podemos decir con exactitud cual es el beta porque tenemos sesgo, podriamos crear algunos "bounds" de cual es el valor del beta
  - Su enfoque intenta formalizar la intuición que motiva las pre-trend test: que las tendencias contrafactuales posteriores al tratamiento no pueden ser "demasiado diferentes" de las tendencias previas.

Partial Identification

- Ye, Keele, Hasegawa y Small (2021) consideran un enfoque alternativo de identificación parcial en el que hay dos grupos de control cuyas tendencias se supone que "encuadran" las del grupo de tratamiento.

Recomendaciones Practicas

- Ver los ES. Freyaldenhoven et al. (2021) tiene recomendaciones de como mostrar los ES.
- Usar la "calculadora" de Roth para enteder el poder de tu pre-trend test.
- Se puede incluir covariates. Siempre de pre-reform.
- Check the "overlap condition" o solo usar "common support".

Recomendaciones Practicas

- Comparar si es que nuestro estimadores cambian mucho con vs. sin variables de control. Idealmente, no cambian.
- El default seria "double-robust" DiD. (Que usa matching + OLS) si es que uno quiere imponer variables de control.
- Reportar el breakdown point the Rambachan and Roth (2022b) $\bar{M}$, ¿Qué tan grande tendría que ser la violación posterior al tratamiento de PT en relación con la violación más grande antes del tratamiento para invalidar una conclusión particular?

Inferencia

- La tercera categoria de papers hablar sobre la inferencia, es decir como estimar CI, SE, etc.
- ¿Que debemos hacer?
  - La regla mas basica de todas: cluster al nivel de la unidad de tratamiento.
  - Munis o personas?
  - Personas o grupo demografico?

Introduction

Staggered Rollout

Bacon Decomp

Resumiendo

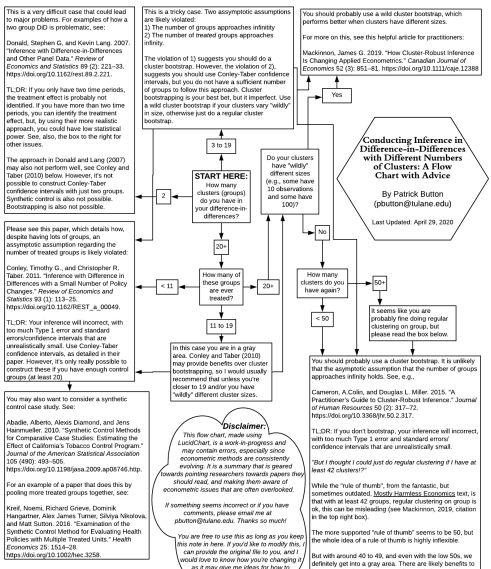
Los nuevos Estimadores

Relaxing Parallel Trends

Inferencia

Otros topicos

Inferencia



¿Que nuevas tendencias hay respecto a la inferencia?

- Leer Section 5 de "*What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature*". Jonathan Roth, Alyssa Bilinski and John Poe (2023).
- Mi recomendacion: Fisher Randomization Tests (FRTs), otherwise known as permutation test or randomization inference
  - Paso 1: correr la regresion principal y guardar el t-statistic (puede ser otras estadisticas).
  - Paso 2: cambiar el D de una manera: Por ejemplo randomizar quien recibe el tratamiento y cuando comienza el tratamiento
  - Paso 3: correr la regresion con este nuevo D, y guardar el t-statistic
  - Paso 4: Repetir Paso 2-3 unas 1,000 veces.

Fisher Randomization Tests (FRTs)

- Paso 1: correr la regresion principal y guardar el t-statistic (puede ser otras estadisticas).
- Paso 2: cambiar el D de una manera: Por ejemplo randomizar quien recibe el tratamiento y cuando comienza el tratamiento
- Paso 3: correr la regresion con este nuevo D, y guardar el t-statistic
- Paso 4: Repetir Paso 2-3 unas 1,000 veces.
- Paso 5: comparar la t-stat original con las 1,000 y reportar: Que porcentage de t-stats la original esta "por encima" de? (Idealmente mayor del 95%, osea esta en el top 5%).
- Comandos: regsave, ritest,

Fisher Randomization Tests (FRTs)

```
generate rannum = uniform()
sort rannum
generate grp = .
replace grp = 0 in 1/50
replace grp = 1 in 51/100
replace grp = 2 in 101/150
replace grp = 3 in 151/200
```

Fisher Randomization Tests (FRTs)

```
generate rannum = uniform()
egen grp2 = cut(rannum), group(4)
```

Recomendaciones Practicas

- Fisher Randomization test
- Simulaciones y probar los distintos tipos de inferencia
- ¿Como hacemos simulaciones?
  - Con tu data crea Ys a traves de tu tratamiento, y efectos fijos + un error que uno crea.
  - Ejemplo: Appendix B (Tello Trillo,2021)

TABLE B1 Monte Carlo simulations

Method	% Rejection of null when p value is .05, $H_o : \beta_{dd} = 0$
OLS	0.098
Robust	0.178
Cluster	0.642
Wild Bootstrap	0.001
Modified block bootstrap	0.055

Note: These results come from 10,000 repetitions using 17 cluster groups.

Otras extension de DiD

- El ultimo set de literatura de DiD trata sobre extensiones del DiD.
- Tal vez no estemos interesados en ATT, sino en la distribucion de los Y
- Athey and Imbens (2006) propose the Changes-in-Changes model.
  - The key assumption is that the mapping between quantiles of Y for the treated and comparison groups remains stable over time – e.g., if the 30th percentile of the outcome for the treated group was the 70th percentile for the comparison group prior to treatment, this relationship would have been preserved in the second period if treatment had not occurred.
- Bonhomme and Sauder (2011); Callaway and Li (2019); Roth and Sant'Anna (2022)

Extensiones de DiD

- Treatment on/off + Y depende de donde viene, y no solo del D y t-1
- Viviano and Bradic (2021) y Robins (1986) tienen algunas propuestas.
- Relajando SUTVA:
  - Spillover effects may be important in many economic applications, such as when policy in one area affects neighboring areas, or when individuals are connected in a network.
- Butts (2021): extending the framework of Callaway and Sant'Anna (2021) to allow for local spatial spillovers.
- Huber and Steinmayr (2021) also consider extensions to allow for spillover effects.

Extensiones de DiD

- Heterogeneity of effects by sub-samples of groups. Por ejemplo: how does the average treatment effect of a training program on earnings vary according to the age of its participants?
- Abadie (2005) provides re-weighting methods to tackle these types of questions using linear approximations.
- However, recent research has shown that data-adaptive/machine-learning procedures can be used to more flexibly estimate treatment effect heterogeneity in the context of RCTs or cross-sectional observa- tional studies with unconfoundedness (e.g., Lee, Okui and Whang, 2017; Wager and Athey, 2018; Chernozhukov, Demirer, Duflo and Fernández-Val, 2020).

Resumen de algunas recomendaciones

Table 1: A Checklist for DiD Practitioners

– **Is everyone treated at the same time?**

If yes, and panel is balanced, estimation with TWFE specifications such as (5) or (7) yield easily interpretable estimates.

If no, consider using a "heterogeneity-robust" estimator for staggered treatment timing as described in Section 3. The appropriate estimator will depend on whether treatment turns on/off and which parallel trends assumption you're willing to impose. Use TWFE only if you're confident in treatment effect homogeneity.

Resumen de algunas recomendaciones

– **Are you sure about the validity of the parallel trends assumption?**

If yes, explain why, including a justification for your choice of functional form. If the justification is (quasi-)random treatment timing, consider using a more efficient estimator as discussed in Section 6.

If no, consider the following steps:

1. If parallel trends would be more plausible conditional on covariates, consider a method that conditions on covariates, as described in Section 4.2.
2. Assess the plausibility of the parallel trends assumption by constructing an event-study plot. If there is a common treatment date and you're using an unconditional parallel trends assumption, plot the coefficients from a specification like (16). If not, then see Section 4.3 for recommendations on event-plot construction.
3. Accompany the event-study plot with diagnostics of the power of the pre-test against relevant alternatives and/or non-inferiority tests, as described in Section 4.4.1.
4. Report formal sensitivity analyses that describe the robustness of the conclusions to potential violations of parallel trends, as described in Section 4.5.

Resumen de algunas recomendaciones

– **Do you have a large number of treated and untreated clusters sampled from a super-population?**

If yes, then use cluster-robust methods at the cluster level. A good rule of thumb is to cluster at the level at which treatment is independently assigned (e.g. at the state level when policy is determined at the state level); see Section 5.2.

If you have a small number of treated clusters, consider using one of the alternative inference methods described in Section 5.1.

If you can't imagine the super-population, consider a design-based justification for inference instead, as discussed in Section 5.2.

Commands

Table 2: Statistical Packages for Recent DiD Methods

Package	Software	Description
Heterogeneity Robust Estimators for Staggered Treatment Timing		
did, csdid	R, Stata	Implements Callaway and Sant'Anna (2021)
did2s	R, Stata	Implements Gardner (2021), Borusyak et al. (2021), Sun and Abraham (2021), Callaway and Sant'Anna (2021), Roth and Sant'Anna (2021)
didimputation, did_imputation	R, Stata	Implements Borusyak et al. (2021)
DIDmultiplegt, did_multiplegt	R, Stata	Implements de Chaisemartin and D'Haultfœuille (2020)
eventstudyinteract	Stata	Implements Sun and Abraham (2021)
flexpaneldid	Stata	Implements Dettmann (2020), based on Heckman et al. (1998)
fixest	R	Implements Sun and Abraham (2021)
stackedev	Stata	Implements stacking approach in Cengiz et al. (2019)
staggered	R	Implements Roth and Sant'Anna (2021), Callaway and Sant'Anna (2021), and Sun and Abraham (2021)
xtevent	Stata	Implements Freyaldenhoven et al. (2019)
DiD with Covariates		
Package	Software	Description
DRDID, drdid	R, Stata	Implements Sant'Anna and Zhao (2020)
Diagnostics for TWFE with Staggered Timing		
Package	Software	Description
bacondecomp, ddtiming	R, Stata	Diagnostics from Goodman-Bacon (2021)
TwoWayFEWeights	R, Stata	Diagnostics from de Chaisemartin and D'Haultfœuille (2020)
Diagnostic / Sensitivity for Violations of Parallel Trends		
Package	Software	Description
honestDiD	R, Stata	Implements Rambachan and Roth (2022b)
pretrends	R	Diagnostics from Roth (2022)

Note: This table lists R and Stata packages for recent DiD methods, and is based on Asjad Naqvi's repository at https://asjadnaqvi.github.io/DiD/. Several of the packages listed under "Heterogeneity Robust Estimators" also accommodate covariates.

Conclusion

- DiD es un research design muy versatil
- La literature se esta preocupando en los efectos heterogenous de tiempo.
- Saber que “comparaciones” esta haciendo el DiD es lo primordial.

Keep in touch!

- Gracias por tenerme!
- Email: Sebastian.tello@viriginia.edu
- Web: sebastiantellotrillo.com
- Web de la clase y pw: udep2023
- twitter: @dsebastiantello
- instagram: @onlyoneacademic
- podcast: The Hidden Curriculum
- Mas informacion de programas y como ingresar: EconThaki