HALF TIME REVIEW

William Lowe

Hertie School

11th October 2020

Half time review

- → Where have we been?
- → Where are we going?

Week 1

In principle, we can think about causation and causal effects in a graphical or a potential outcomes terms

→ Columns / variables (graphs of mechanisms)
→ Rows / cases / individuals (potential outcomes)

In practice, the choice is a question of research assumptions

→ Stable, autonomous causal *mechanisms*
→ Population-specific causal *effects* (ATEs, ATTs etc. are always relative to a population)

and of worries (these are slightly stereotypical)

→ Mechanisms have homogeneity of operation as the rule, and individual heterogeneity as the exception
→ Individual heterogeneity of effects is the rule, and homogeneity of operation the exception

Week 1

- → When mechanical approaches need heterogeneity they condition on something else
- → When PO approaches need uniformity or other extra structure they assume it with constraints, e.g. equality of PO or 'monotonicity'

Machine Learning, e.g function approximation methods are an interesting compromise

- → Condition on, well, all any anything you can give them
- → Assume that generates all the heterogeneity in estimates we need
- → Can, in principle, estimate *all the interactions*

Week 2

We talked about experiments, and how they go wrong

→ through individual heterogeneity (in compliance, dropout, etc)

The bigger picture:

→ The *experimental* ideal gets awkwardly *observational* fast

In theory, experiments are for effects, not mechanisms

→ But maybe we don't care about the population they were performed on
→ We just want to transport the results to another one (via interactions, or reweighting)
→ This is a more mechanical concern... (and instrumental variable estimation requires more mechanical assumptions)

Week 3

Linear regression is a very mechanical tool. Implicitly

→ assumes *homogeneous* effects identified by coefficient estimates

We can *retrofit* causal interpretations (sometimes) but the regression does not 'know' which is the treatment variable, and has other troubles…

There are *no* reliable data-driven ways to decide whether a variable belongs in a regression

→ You've got to know more about the mechanisms (Cinelli et al., 2020)

Week 3: Weight control

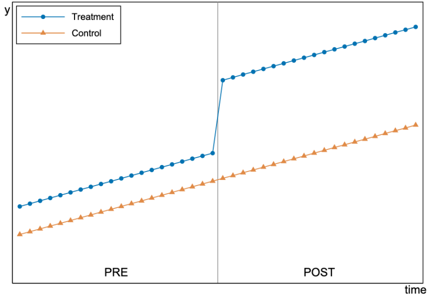
Heterogeneity leads to (perhaps unintuitive) weighting interpretations

- → Aronow and Samii (2016) on the effects of democracy on FDI
- → Goodman-Bacon (2018) and the interpretation of diff-in-diff models with variations in treatment timing

Let's step briefly into Goodman-Bacon's results

Interlude: Diff in diff

Our old friend the difference in differences design, e.g. Card and Krueger (2000)



If we're willing to assume *parallel trends* then

$$\left(\mathrm{E}[Y_{\text{post}}^{T=1}] - \mathrm{E}[Y_{\text{pre}}^{T=1}]\right) \;-\; \left(\mathrm{E}[Y_{\text{post}}^{T=0}] - \mathrm{E}[Y_{\text{pre}}^{T=0}]\right)$$

is the ATT

Interlude: Diff in diff

Conveniently we can think of this as a simple regression with an interaction

$$Y_{it} = \alpha + T\beta_T + (\text{post})\gamma_T + (T \times \text{post})\beta^{DD} + \epsilon_{it}$$

which in turn boils down to a two-way fixed effects model

$$Y_{it} = \alpha_i + \gamma_t + T\beta^{DD} + \epsilon_{it}$$

So far, so straightforward.

Interlude: Diff in diff

Conveniently we can think of this as a simple regression with an interaction

$$Y_{it} = \alpha + T\beta_T + (\text{post})\gamma_T + (T \times \text{post})\beta^{DD} + \epsilon_{it}$$

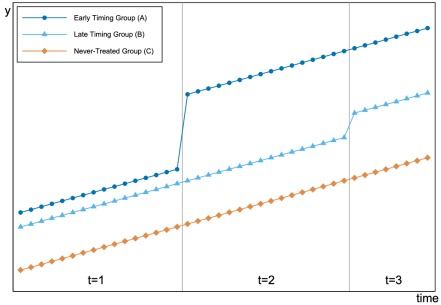
which in turn boils down to a two-way fixed effects model

$$Y_{it} = \alpha_i + \gamma_t + T\beta^{DD} + \epsilon_{it}$$

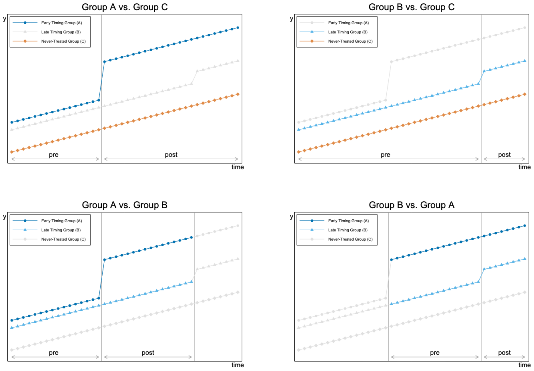
So far, so straightforward.

Now the interpretation of β^{DD} is less straightforward

But what if



Interlude: Diff in diff



It's a *mixture* of all these possible diff-in-diff designs, weighted by

→ treatment variance (like regular regression)

→ sample size in each timing region

Interlude: Diff in diff

Consequences:

- → Homogenous treatment effects? Then β^{DD} is the ATE
- → Treatment effects heterogenous across cases? Variance weighted ATE (i.e. not the ATE)
- → Treatment effects heterogenous over time? Bias…

Solutions?

- → Matching / weighting (Imai & Kim, 2019)
- → Event studies (though see Borusyak & Jaravel, 2016)

I found these slides (link) so helpful that I stole the figures.

The author's slides are also good: (link)

For practical purposes R has a bacondecomp package. For solutions, the PanelMatch and wfe package offer matching estimators (implementing the Song, Wang, and Imai references.)

Week 4

A first introduction to ML, at least as function approximation

- (Nearly infinitely) flexible functional forms: interact all the things!
- Suited to and expecting high dimensional (many variable) data sources

New possibilities

- Condition on all the things

New problems

- Conditioning on all the things…
- Lack of bias brings variance, unless properly regularized
- Like linear regression, treatment is not special

Week 4

These are unfamiliar tools

- → Causal effects get *very* conditional: both on the treatment variable (if continuous) and on potentially every other covariate
- → Marginal effects are tricky to compute and tricky to interpret (averages over high dimensional Z can be weird)
- → A useful reminder of persistent the linear / difference thinking among the causal effect devotees

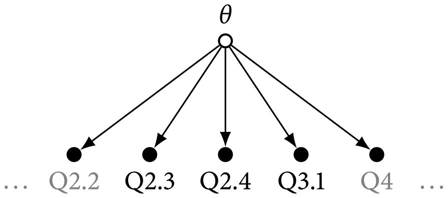
We also saw the promise and problems of very high dimensional problems:

- → Ideological preference estimation is not normally thought of a causal inference problem
- → As measurement, it should be *easier*

Time for a micro-interlude on measurement…

Interlude: Measurement

Measurement models, of the kind discussed in Broockman (2016) all look like this:



This model works best when

- → each arrow is a causal mechanism
- → each mechanism is relatively autonomous
- → there are no parents of each response that systematically affect it in interaction with θ

Example:

- → Responses to Q2.3 vary by country
- → or worse, country confounds θ and response

a.k.a. differential item functioning (DIF)

Week 5

With modern tools and data sources we can 'go big' on

- → variables: lots of covariates, lots of interactions, arbitrary non-linearity
- → cases: 'big data'

In principle, more cases bring nothing new but some good to causal inference

- → more power to find interactions
- → more power to estimate the causes annd effects of rare events

In practice:

- → Big data data sets are often automatically and indiscriminately collected
- → or for purposes that are unrelated to causal inference purposes, e.g. legal compliance, business process optimization, etc.
- → ...huge selection bias possibilities

Week 5: Big data

Intuition: selection bias can be expected when the thing being measured affects whether you get to measure it

- → Much more on this after the break

Huge amount of statistical theory is based on getting, or keeping, a *random sample*

- → The basis for the standard error calculation that include sample size, but not population size

Often forgotten…

Interlude: Data quality

Meng (2018) starts with a question relevant to all of us

> *Which one should we trust more, a 5% survey sample or an 80% administrative dataset?*

Interlude: Data quality

Meng (2018) starts with a question relevant to all of us

Which one should we trust more, a 5% survey sample or an 80% administrative dataset?

To see this, we first let $\rho_{R,G} = \mathrm{Corr}_J(R_J, G_J)$ be the (population) correlation between R_J and G_J, $f = \mathrm{E}_J(R_J) = n/N$ be the sampling rate, and σ_G be the standard deviation of G_J, all defined according to the uniform distribution of J. Then, using the fact that the variance of the binary R_J is $\mathrm{V}_J(R_J) = f(1-f)$, we have from (2.2) that

$$\overline{G}_n - \overline{G}_N = \underbrace{\rho_{R,G}}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{\text{Data Quantity}} \times \underbrace{\sigma_G}_{\text{Problem Difficulty}}. \tag{2.3}$$

The expected value for ρ under most random sampling schemes is $1/(n-1)$

Interlude: Data quality

Perhaps more surprisingly

- → We can offset non-random sampling by increasing sample size
- → But it gets out of control really quickly (details in the paper)

Meng states this as a survey theory result, but it's of key relevance to causal inference:

- → Most business and administrative data is not randomly selected in this sense
- → Most often selected on a post-treatment variable
- → It's (mathematically) hard to offset this by collecting more data…

Where are we going

Further into policy

- → Collider bias as far as the eye can see
- → Fairness, algorithmic and otherwise
- → Mechanisms and effects together at last (in mediation)

Onwards, into the inevitable lockdown…

References

Aronow, P. M. & Samii, C. (2016). 'Does regression produce representative estimates of causal effects?' *American Journal of Political Science*, *60*(1), 250–267.

Borusyak, K. & Jaravel, X. (2016). 'Revisiting event study designs'. *SSRN Electronic Journal*.

Broockman, D. E. (2016). 'Approaches to studying policy representation'. *Legislative Studies Quarterly*, *41*(1), 181–215.

Card, D. & Krueger, A. B. (2000). 'Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania: Reply'. *American Economic Review*, *90*(5), 1397–1420.

Cinelli, C., Forney, A. & Pearl, J. (2020). 'A crash course in good and bad controls'. 10.

Goodman-Bacon, A. (2018, September). *Difference-in-differences with variation in treatment timing* (w25018). National Bureau of Economic Research. Cambridge, MA.

References

Imai, K. & Kim, I. S. (2019). 'When should we use unit fixed effects regression models for causal inference with longitudinal data?' *American Journal of Political Science*, *63*(2), 467–490.

Meng, X.-L. (2018). 'Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election'. *The Annals of Applied Statistics*, *12*(2), 685–726.