Proxy Controls and Panel Data

Ben Deaner (MIT)

Econometric Society World Congress 7/20/20

Motivation

- Confounding is the key challenge of empirical economics.
- Ideally one can identify causal objects by controlling for observables that satisfy an 'unconfoundedness' assumption.
- In practice 'perfect controls' are often unavailable, but one has informative proxies for some ideal perfect controls.

  $\rightarrow$ e.g., test scores as a proxy for academic ability.
- Treating these 'proxy controls' like conventional perfect controls leads to asymptotic bias (measurement error).

The contribution (1/2)

- I provide new results for nonparametric identification of the conditional average structural function (CASF) with only proxy controls.
- The identification results suggest an estimator based on a conditional moment restriction.
- Essentially, one splits the proxies into two groups and instruments for one with the other.
  - $\rightarrow$ Easy to implement. No need for numerical optimization or simulation.
  - $\rightarrow$ I derive simple convergence rates under intuitive assumptions.
- All nonparametric, no separability assumptions, no sufficient statistics, no shape restrictions.

The contribution (2/2)

- My analysis applies to cross-sectional models, but is particularly applicable to panels with fixed T.
- In panels past observations act as proxies for unobserved individual heterogeneity.
- Analogous to classical panel models like fixed-effects.
- Restrictions on the panel structure imply key conditional independence assumptions.
- I present two empirical applications
  - → One is based on cross-sectional variation (impact of grade retention)
  - → The other is based on the panel structure (demand for food)

Related literature (1/3)

- Miao et al (2018 Biometrika) prove identification of the average structural function under very similar assumptions.
- Differences:
  - → I identify the CASF, establish 'well-posedness', and relate the high-level assumptions to panel structure.
  - → They do not provide an estimator. I do so and I establish consistency and a convergence rate.

Related literature (2/3)

- Hu and Schennach (2008 Econometrica), consider non-separable measurement error in nonparametric models.
- Rokkanen (2015 working paper) applies this to regression discontinuity with proxy controls.
- These papers have some similar assumptions to mine: conditional independence restrictions and statistical completeness.
- Differences:
  - → They require a normalization like mean or median unbiasedness of error-prone measurements. And estimation is complicated.
  - → My problem is easier than Hu and Schennach's because the measurement error is only in controls (not treatments).
  - → I do not back out any distributions involving latent factors and my method cannot be used for this.

Related literature (3/3)

- Arellano and Bonhomme (2016 The Econometrics Journal), use a factor analytic approach related to Hu and Schennach in panel settings.
- Differences:
  - → They identify distributions of latent factors (I do not).
  - → They present a complicated MCMC-based estimator and require some separability in the factor structure.
- There are myriad papers related to my empirical examples (the impact of grade-retention on future success and analysis of the demand for food). See the paper for references.

What are proxy controls?

- I define them in contrast to 'perfect controls'.
- Conditioning on perfect controls, treatment assignments and potential outcomes are independent (unconfoundedness).
- Proxy controls are informative about some unobserved perfect controls, but are not themselves perfect controls.
- Hence 'proxy'. They act as proxies for the unobserved perfect controls.

"Okay, but give me an example."

- The impact of being held back a grade on future test scores:
  - → Academic merit at time of being held back is the key confounding factor.
  - → Cannot observe academic merit, but test scores prior to being held back may be good proxies.
- The demand for food:
  - → Use past expenditure on non-durables to proxy for household consumption preferences.

Setting up the problem

- Non-parametric and non-separable causal model:

$$Y = y_0(X, U)$$

- X are observed treatment assignments, Y is observed outcome. U is unobserved heterogeneity in potential outcomes (need not have finite dimension).
- $y_0(x, u)$ is the potential outcome from counterfactual treatment level x for an agent with heterogeneity u.
- Could be cross-sectional or a model for outcomes at time t in panel settings (more on this later).

Continuing to set up the problem

- The conditional average structural function (CASF) is defined by:

$$E[y_0(x_1, U)|X = x_2]$$

- This is the average potential outcome from treatment x_1 for agents whose observed assigned treatment is x_2.
- Key counterfactual objects of interest can be written in terms of the above. e.g., the effect of treatment on the treated.

What are perfect controls?

- Suppose there is some vector of covariates W^* so that conditional on this vector, treatment assignments and heterogeneity are independent:

$$U \perp\!\!\!\perp X | W^*$$

- Then W^* is a 'perfect control'.
- If W^* has satisfies something like a full-support assumption and some conditional means exist and are finite then the CASF is identified if W^* is observed.

Once more, what are proxy controls?

- Proxy controls proxy for some unobserved perfect control W^* but need not be perfect controls.
- Split the available proxy controls into two vectors V and Z.
- Z is essentially used to instrument for V.

The identifying assumptions

- W^* is a perfect control and a full support assumption holds.
- Variation in Z after conditioning on W^* and X is independent of U. Formally, $U \perp\!\!\!\perp Z | (W^*, X)$.
- All association between V and (Z, X) is explained by W^*. Formally $V \perp\!\!\!\perp (X, Z) | W^*$.
- Both V and Z must each be sufficiently informative about W^*. Given in terms of statistical completeness.
- Also require a technical regularity condition in terms of generalized Fourier coefficients.

Conditional independence

- The assumption $V \perp\!\!\!\perp (X, Z) | W^*$ holds if V and (Z, X) are each functions of W^* and some orthogonal noise.
  - $\rightarrow$ Grade retention example: if Z and V are sets of test scores which reflect academic merit W^* and treatments X (whether child must repeat a grade) is determined by other measures of W^* like grades and behavior.
- The proxies V could be very strongly correlated with treatments X and proxies Z, but all of the correlation must be explained by mutual association with latent perfect controls W^*.
- If some perfect controls (components of W^*) are observed they should be included in both V and Z.

Informativeness

- V and Z must be sufficiently informative about W^*. Analogous to instrumental relevance, need each of V and Z to be relevant for W^*.
- The informativeness assumptions are in terms of statistical completeness (more precisely, L_2-completeness).

$$E[\delta(W^*)|Z, X = x] = 0 \iff \delta = 0$$

$$E[\delta(W^*)|V, X = x] = 0 \iff \delta = 0$$

- This is a nonparametric analogue of the rank condition in linear IV (see Newey and Powell (2003) Econometrica).
- Order condition: should ensure V and Z have as many components as W^*. Strictly speaking, not necessary for completeness.

Regularity conditions

- Also impose a technical regularity condition given in terms of square-summability of a sequence involving generalized Fourier coefficients.
- Similar assumptions are made in the NPIV context, e.g., Darolles, Florens & Renault (2011 Econometrica).

Time to identify the CASF!

- I give two different characterizations of the CASF in terms of observables. Either could motivate an estimator, but I focus on the following one.
- Under the assumptions there is a γ that solves the following conditional moment restriction (that involves only observables):

$$E[Y - \gamma(X, V)|X, Z] = 0$$

- And for any such a γ one can recover the CASF as follows:

$$E[y_0(x_1, U)|X = x_2] = E\left[\gamma(x_1, V)|X = x_2\right]$$

And the problem is well-posed

- Many conditional moment restriction problems are 'ill-posed' (e.g., NPIV estimation of the structural function).
- Well-posedness is very important, it allows one to derive fast and simple convergence rates. It is also important for robustness to misspecification, see Deaner (2019 R&R Econometrica).

Panel data

- Let's add time-subscripts:

$$Y_t = y_{0,t}(X_t, U_t)$$

- W^*, V and Z will not have subscripts, could contain variables from multiple periods, but they are individual-specific.
- Indeed, in the discussion to follow I consider V and Z made up of treatments in earlier periods.
- Identify the CASF at period T, which is $E[y_{0,T}(x_1, U_T)|X_T = x_2]$.

The panel structure can help

- If V and Z are each made up observations at different periods, then $V \perp\!\!\!\perp (Z, X_T) | W^*$ is an assumption on the serial dependence structure.
- In this talk I'll discuss one case in which V and Z are composed of past treatment assignments.

Pre-determination and Markov dependence

- Suppose that conditional on some latent variables $\tilde{W}^*$, the following conditional independence restriction holds:

$$U_t \perp\!\!\!\perp (X_1, ..., X_t) | \tilde{W}^*$$

- This is analogous to pre-determination in linear fixed effects models. Past treatments do not effect shocks today or in future.
- Conditional on the latent variables $\tilde{W}^*$, the treatments satisfy a first-order Markov dependence structure. Formally:

$$(X_1, ..., X_{t-1}) \perp\!\!\!\perp (X_{t+1}, ..., X_T) | (\tilde{W}^*, X_t)$$

Now let's form proxy controls

- Let V consist of treatments prior to $\lfloor T/2 \rfloor$ inclusive, let Z consist of treatments from $\lfloor T/2 \rfloor$ to $T-1$ inclusive. Take $W^* = (\tilde{W}^*, X_{\lfloor T/2 \rfloor})$.
  $\rightarrow$ Note that $X_{\lfloor T/2 \rfloor}$ is an observed perfect control and is included in both Z and V.
- Then our conditional independence assumptions hold:
  $U_T \perp\!\!\!\perp (X_T, Z) | W^*$, $V \perp\!\!\!\perp (X_T, Z) | W^*$.

Are the proxies informative enough?

- Recall that each of V and Z must be sufficiently informative about W^* (relevant instruments).
- If $\tilde{W}^*$ explains some of the confounding in each period (say, it is time-invariant), then by construction it will be associated with treatments in each period, and thus with each component of V and Z.
- The order condition requires W^* have fewer components than V and Z. With scalar treatments, $\tilde{W}^*$ must have fewer components than $\lfloor T/2 \rfloor - 1$.

  $\rightarrow$ Bigger T means the informativeness assumption is ore plausible.

Intuition

- Find $\hat{\gamma}$ that approximately satisfies an empirical analogue of:

$$E[Y - \gamma(X, V)|X, Z] = 0$$

- Recall the CASF equals $E\left[\gamma(x_1, V)|X = x_2\right]$. Estimate the CASF using an empirical analogue of:

$$E_V\left[\hat{\gamma}(x_1, V)|X = x_2\right]$$

- The procedure looks like NPIV, particularly penalized sieve minimum-distance.

The first step

- Choose a vector of basis functions ϕ_n defined on support of (X, V) with dimension that grows with n.
- Use your favorite nonparametric regression method (e.g, series, ridge, lasso, local-linear) to estimate:

$$\hat{g}_i \approx E[Y|X_i, Z_i]$$
$$\hat{\pi}_{n,i} \approx E[\phi_n(X_i, V_i)|X_i, Z_i]$$
$$\hat{\alpha}_n(x_1, x_2) \approx E[\phi_n(x_1, V_i)|X_i = x_2]$$

The second step

- Let λ_n be a positive scalar penalty parameter. Minimize:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}_i - \hat{\pi}'_{n,i}\theta\right)^2 + \lambda_n||\theta||_2^2$$

- Let $\hat{\theta}$ minimize the above. The final estimate is:

$$E[y_0(x_1, U)|X = x_2] \approx \hat{\alpha}_n(x_1, x_2)'\hat{\theta}$$

- In my applications I use series ridge regression for first stage so the estimator has a simple formula.

Consistency and a convergence rate

- In the discrete treatment case one can achieve rate $n^{-\frac{s}{2s+\dim(Z)+\dim(V)}}$ where s is the smoothness class of some reduced-form functions.
- Paper contains very general rate results (continuous treatments, sub-optimal penalty parameters and basis choices).
- All derived under primitive conditions of the kind typical in non-parametric regression literature.
- Well-posedness plays a key role. No 'sieve-measure of ill posedness'.

The first application: grade retention

- Based on Fruehwirth, Navarro & Takahashi (2016 Journal of Labor Economics). I use their cleaned data which is from the ECLS-K panel study (1998-1999 kindergarten year).
- Goal is to examine the causal effect of grade retention on the cognitive outcomes of children as measured by tests in reading, math and science at age ≈ 11.
- Treatments are retention in kindergarten, 'early' (in first or second grade) and 'late' (in third or fourth grade). Same as Fruehwirth (2016).
- Fruehwirth (2016) employ a factor model with academic ability a latent factor and use test scores to back out factor loadings.
  $\rightarrow$ Our approach allows for a more flexible factor structure and simpler estimation method.

Some results!

TABLE I

EFFECTS OF GRADE RETENTION ON COGNITIVE PERFORMANCE

(a) Reading Ability

$n = 1998$	**Observed retention status:**			
Difference from non-retention:	Not retained	Retained kindergarten	Retained early	Retained late
Retained kindergarten	-0.07	0.07	-0.05	0.14
	(0.12)	(0.14)	(0.30)	(0.27)
Retained early	-0.07	0.11	0.01	-0.05
	(0.13)	(0.26)	(0.15)	(0.24)
Retained late	-0.33	-0.25	-0.32	0.15
	(0.14)	(0.43)	(0.43)	(0.16)

(b) Math Ability

$n = 1999$	**Observed retention status:**			
Difference from non-retention:	Not retained	Retained kindergarten	Retained early	Retained late
Retained kindergarten	-0.12	0.13	0.06	0.15
	(0.14)	(0.19)	(0.29)	(0.24)
Retained early	-0.17	0.12	0.11	-0.02
	(0.16)	(0.34)	(0.17)	(0.30)
Retained late	-0.43	-1.30	-1.38	0.02
	(0.22)	(0.67)	(0.75)	(0.15)

The second application: demand for food

- To demonstrate the application to panel models I apply the method to estimation of food-demand counterfactuals.
- Structural Engel curves: budget share on food (at home) under exogenously chosen total expenditure. Also, average change in the budget share of food from an exogenous 10% increase in total expenditure, broken down by current total expenditure.
- Structural Engel curves for food using 10 periods of data from the Panel Study of Income Dynamics (PSID).

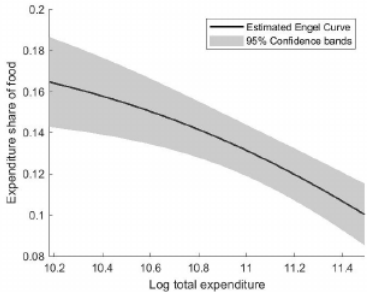
What proxies to use?

- Assume the Markov treatment assignment and predetermination case discussed earlier. The role of assets suggests that allowing for (at least) Markov dependence in treatments is important.
- $\tilde{W}^*$ captures underlying household consumption preferences.
- Assume shocks to preferences non-durables are only related to the history of total expenditure through the mutual association with $\tilde{W}^*$.
- Assume that total expenditure in the past and in the future are only related through the expenditure today and the consumption preferences.

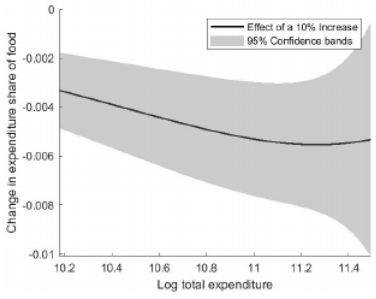
Some more results!

Figure 1: Demand for Food

(a) Average Engel Curve for Food



(b) Change in Food Demand from 10% Total Expenditure Rise



What is left to do?

- Still working on primitive conditions to justify bootstrap-based inference.
- Future work will develop a Neyman-orthogonal (locally robust) version of the estimator and consider a growing number of proxies (also covers $T \to \infty$).