Chapter 10: Nonlinear Models

Peter Hull

Applied Econometrics II
Brown University
Spring 2024

Outline

1. Limited Dependent Variables

2. Control Functions

3. Demand Estimation

Why So Linear?

The vast majority of this class has focused on linear regression/IV

- Why? Recall the CEF approximation result in Chapter 1
- Results from Angrist (1998)/Imbens & Angrist (1994)/etc let us causally interpret OLS/IV estimands when CEFs aren't actually linear

Still, you might sometimes want to fit nonlinear models

- A leading justification: Y_i has limited support (e.g. binary), and you want predictions $\hat{Y}_i$ to respect this (e.g. be in $[0, 1]$)
- Another reason is *extrapolation*: nonlinear models implicitly structure heterogeneity in causal effects (differently than OLS)
- Let's consider these in turn...

LDPs in Saturated Models

Suppose you run an RCT with a binary treatment D_i. We know regression of Y_i on D_i identifies $E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = E[Y_i(1) - Y_i(0)]$

- What changes if Y_i is binary? Nothing.

Suppose the true model is a Probit:

$$\begin{aligned} Y_i &= \mathbf{1}[Y_i^* \geq 0] \\ Y_i^* &= \alpha + \beta D_i - \varepsilon_i \\ \varepsilon_i \,|D_i &\sim \mathrm{N}(0, \sigma^2) \end{aligned}$$

The CEF implied by this model is linear — regression fits it perfectly

$$\begin{aligned} E[Y_i \mid D_i] = Pr(Y_i^* \geq 0 \mid D_i) &= \Phi\left(\frac{\alpha + \beta D_i}{\sigma}\right) \\ &= \underbrace{\Phi\left(\frac{\alpha}{\sigma}\right)}_{\text{constant}} + \underbrace{\left\{\Phi\left(\frac{\alpha+\beta}{\sigma}\right) - \Phi\left(\frac{\alpha}{\sigma}\right)\right\}}_{\text{slope coefficient}} D_i \end{aligned}$$

LDPs in Saturated Models — in the Sample

It follows that both regression and Probit can both identify the ATE

- In fact, OLS and Probit *estimates* of $E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$ will end up being the same in any given sample

Let's test this claim in the Oregon Health Insurance Experiment data:

```
. reg health_notpoor_12m treatment, r

Linear regression                               Number of obs     =     23,397
                                                F(1, 23395)       =      52.84
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0023
                                                Root MSE          =     .33019

------------------------------------------------------------------------------
             |               Robust
health_n~12m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |   .0313635   .0043147     7.27   0.000     .0229063    .0398207
       _cons |   .8595596   .0032036   268.31   0.000     .8532803     .865839
------------------------------------------------------------------------------
```

Getting access to Medicaid increases $Y_i = \mathbf{1}[\text{Not poor health}_i]$ by 3.1pp

LDPs in Saturated Models — in the Sample

It follows that both regression and Probit can both identify the ATE

- In fact, OLS and Probit *estimates* of $E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$ will end up being the same in any given sample

Let's test this claim in the Oregon Health Insurance Experiment data:

```
. probit health_notpoor_12m treatment, r

Iteration 0:   log pseudolikelihood =  -8808.235
Iteration 1:   log pseudolikelihood = -8781.8488
Iteration 2:   log pseudolikelihood = -8781.8333
Iteration 3:   log pseudolikelihood = -8781.8333

Probit regression                               Number of obs     =     23,397
                                                Wald chi2(1)      =      52.63
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -8781.8333               Pseudo R2         =     0.0030

------------------------------------------------------------------------------------
                   |               Robust
health_notpoor_12m |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
         treatment |   .1531095   .0211048     7.25   0.000     .1117449    .1944741
             _cons |   1.078343   .0143624    75.08   0.000     1.050193    1.106493
------------------------------------------------------------------------------------
```

Wait ... what's going on here?

Coefficients vs. Marginal Effects

The (normalized) probit model is $Y_i = \mathbf{1}\left[\frac{\alpha}{\sigma} + \frac{\beta}{\sigma} D_i > \nu_i\right]$, $\nu_i \mid D_i \sim \mathrm{N}(0,1)$

- Stata reports normalized coefficients, α/σ and β/σ
- These won't give the size of the ATE until we feed them back into $\Phi(\cdot)$ (though β/σ will have the right sign)

The *mfx* postestimation command will give the Probit "marginal effects" (here ATE) complete with SEs; now it matches with OLS!

```
. mfx

Marginal effects after probit
      y  = Pr(health_notpoor_12m) (predict)
         =  .87584739
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.	]	X
treatm~t*	.0313635	.00431	7.27	0.000	.022907	.03982	.497243

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Why Does Probit Match OLS?

1) Probit-MLE is a particular weighted nonlinear least squares procedure

$$Y_i = \mathbf{1}[X_i'\gamma \geq \varepsilon], \quad \varepsilon_i \mid X_i \sim F$$

Log likelihood: $\ln \mathscr{L} = \ln \left(\prod_i F(X_i'\gamma)^{Y_i}(1 - F(X_i'\gamma)^{1-Y_i} \right)$

F.O.C.: $$0 = \sum_i \left(\frac{Y_i - F(X_i'\hat{\gamma}^{MLE})}{F(X_i'\hat{\gamma}^{MLE})(1 - F(X_i'\hat{\gamma}^{MLE}))} \right) f(X_i'\hat{\gamma}^{MLE})X_i$$

Compare to WNLS based on $E[Y_i \mid X_i] = F(X_i'\gamma)$ with weights ω_i:

$$\hat{\gamma}^{WNLS} = \arg\min_c \sum_i \omega_i (Y_i - F(X_i'c))^2$$

F.O.C.: $$0 = \sum_i \omega_i (Y_i - F(X_i'\hat{\gamma}^{WNLS}))f(X_i'\hat{\gamma}^{WNLS})X_i$$

MLE=WLNS with inv-var weights $\omega_i = \left(F(X_i'\hat{\gamma}^{MLE})(1 - F(X_i'\hat{\gamma}^{MLE}))\right)^{-1}$

Why Does Probit Match OLS? (cont.)

2) With saturated $X_i = (1, D_i)'$, WNLS can be rewritten as WLS:

$$\begin{aligned}\hat{\gamma}^{WNLS} &= \arg\min_c \sum_i \omega_i (Y_i - F(X_i'c))^2 \\ &= \arg\min_c \sum_i \omega_i (Y_i - F(c_0) - (F(c_0 + c_1) - F(c_0))D_i)^2\end{aligned}$$

with constant $F(\hat{c}_0{}^{WNLS})$ and slope $F(\hat{c}_0{}^{WNLS} + \hat{c}_1{}^{WNLS}) - F(\hat{c}_0{}^{WNLS})$

3) With saturated X_i, the weights in WLS don't matter

- There is only one estimate of $E[Y_i \mid X_i]$, given by group means!

Whose CEF Estimate Is It Anyway?

So long as our X_i specification is "flexible" enough, it doesn't matter if we use OLS or Probit (or Logit, etc.) — regardless of the support of Y_i

- If we use OLS we get differences in the CEF directly; with Probit/etc. we may need to do a bit more work after estimation

What if X_i is not saturated? The story changes

- Probit will always give fitted values in $(0, 1)$, which can be desirable (e.g. for estimating propensity scores); OLS can predict outside $(0, 1)$
- Both OLS and Probit can be understood as *extrapolating* across the support of X_i, which may or may not be desirable

Example: RCT With Overlap Failures

Suppose individuals with baseline covariate $X_i = 1$ are randomized into a treatment $D_i \in \{0,1\}$. Individuals with $X_i = 0$ are untreated

- Q: What do we get from regressing Y_i on D_i, controlling for X_i?
- A: The CATE $E[Y_i(1) - Y_i(0) \mid X_i = 1]$ (recall, e.g., Angrist '98)
- If we take the linear model for $E[Y_i \mid D_i, X_i]$ seriously, this is also our estimate of the (otherwise unidentified) $E[Y_i(1) - Y_i(0) \mid X_i = 0]$

Suppose Y_i is binary and we instead run a Probit on D_i and X_i

- Taking it seriously, we have a model of heterogeneous effects:

$$Y_i(0) = \mathbf{1}[\alpha + \gamma X_i \geq \varepsilon_i], \quad Y_i(1) = \mathbf{1}[\alpha + \beta + \gamma X_i \geq \varepsilon_i]$$

- In particular, $E[Y_i(1) - Y_i(0) \mid X_i = 1] = \Phi(\frac{\alpha+\beta+\gamma}{\sigma}) - \Phi(\frac{\alpha+\gamma}{\sigma})$ (same as w/*reg*) but $E[Y_i(1) - Y_i(0) \mid X_i = 0] = \Phi(\frac{\alpha+\beta}{\sigma}) - \Phi(\frac{\alpha}{\sigma})$
- $E[Y_i(1) - Y_i(0) \mid X_i = 1] \neq E[Y_i(1) - Y_i(0) \mid X_i = 0]$: a feature or bug?

So Long; Thanks For The Poisson

One nonlinear CEF which can be very useful/tractable is *Poisson*:

$$E[Y_i \mid X_i] = \exp(X_i'\beta) \implies \ln(E[Y_i \mid X_i]) = X_i'\beta$$

Developed for count data, $Y_i \in \{0, 1, 2, \dots\}$. But like OLS, Poisson-MLE has a nice robustness property: consistent when the CEF is well-specified

This makes Poisson a nice approach for modeling non-negative Y_i with zeros but a long right tail (e.g. income)

- Less good idea in such settings: OLS with $\ln(1 + Y_i)$ or $arcsinh(Y_i)$ as the outcome (see e.g. Chen and Roth (2023) for why)
- See Wooldridge (2022) for discussion of how Poisson-(Quasi-)MLE works in diff-in-diff (key: parallel trends in growth rates)

Do OLS and NLS MFX Really Differ Much in Practice?

		Right-hand side variable									
		More than two children						Number of children			
	Mean	OLS	Probit		Tobit		OLS	Probit MFX	Tobit MFX	
			Avg effect, full sample	Avg effect on treated	Avg effect, full sample	Avg effect on treated		Avg effect, full sample	Avg effect, full sample	Avg effect on treated
Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Full Sample										
Employment	.528	-.162	-.163	-.162	-	-	-.113	-.114	-	-
	(.499)	(.002)	(.002)	(.002)			(.001)	(.001)		
Hours worked	16.7	-5.92	-	-	-6.56	-5.87	-4.07	-	-4.66	-4.23
	(18.3)	(.074)			(.081)	(.073)	(.047)		(.054)	(.049)
Panel B: Non-white College Attendees over 30, first birth before age 20										
Employment	.832	-.061	-.064	-.070	-	-	-.054	-.048	-	-
	(.374)	(.028)	(.028)	(.031)			(.016)	(.013)		
Hours worked	30.8	-4.69	-	-	-4.97	-4.90	-2.83	-	-3.20	-3.15
	(16.0)	(1.18)			(1.33)	(1.31)	(.645)		(.670)	(.659)

MHE Table 3.4.2. Of course, it is possible to come up with examples where the difference is large (see, e.g., Lewbel et al. (2012))

Outline

1. Limited Dependent Variables✓
2. Control Functions
3. Demand Estimation

Sample Selection Models

One popular use of nonlinear models is the canonical Heckman (1976, 1979) approach to sample selection ("Heckit"). Suppose $Y_i = Y_i^* D_i$,

$$D_i = \mathbf{1}[X_i'\pi > v_i], \quad v_i \mid X_i \sim \mathrm{N}(0,1)$$
$$Y_i^* = X_i'\beta + \varepsilon_i, \qquad \varepsilon_i \mid v_i, X_i \sim \mathrm{N}(\rho v_i, \sigma^2)$$

This is a fully parametric model of (Y_i, D_i, X_i); we can estimate it by MLE

- But Heckman showed there's an easier two-step approach...

Note $E[Y_i \mid D_i = 1, X_i] = X_i'\beta + E[\varepsilon_i \mid D_i = 1, X_i] = X_i'\beta + \rho \frac{\phi(X_i'\pi)}{1-\Phi(X_i'\pi)}$; here $\frac{\phi(z)}{1-\Phi(z)}$ is the *inverse Mills ratio*. Now we can:

1. Estimate π by Probit of D_i on X_i
2. Estimate β by OLS of Y_i on X_i and $\underbrace{\dfrac{\phi(X_i'\hat{\pi})}{1-\Phi(X_i'\hat{\pi})}}_{\text{control function}}$ cond. on $D_i = 1$

Then we get the unselected relationship and mean, $\hat{E}[Y_i^*] = \hat{E}[X_i]'\hat{\beta}$

From Sample Selection to Causal Inference

For $Y_i = (1-D_i)Y_i(0) + D_i Y_i(1)$, we can apply the same Heckit approach to estimate $E[Y_i(1) - Y_i(0)]$ (or any other causal parameter)

- I.e. $Y_i(1) = X_i'\beta_1 + \varepsilon_{1i}$, $Y_i(0) = X_i'\beta_0 + \varepsilon_{0i}$, $(\varepsilon_{1i}, \varepsilon_{0i}) \mid \nu_i \sim \mathrm{N}\ldots$
- After selection-correcting, $\hat{E}[Y_i(1) - Y_i(0)] = \hat{E}[X_i](\hat{\beta}_1 - \hat{\beta}_0)$

This is identification "purely off of functional form"

- Instead of selection-on-observables, we model the dependence of $(Y_i(1), Y_i(0))$ and D_i conditional on X_i ... and take it seriously!
- Of course, we could have picked a different distribution for $(\varepsilon_{1i}, \varepsilon_{0i}, \nu_i)$ and gotten a different estimate; how do we pick?
- More practical issue: $\phi(z)/(1-\Phi(z))$ turns out to be approximately linear in z, so the control function is likely close to collinear with X_i

Control Functions (with Instruments)

Suppose we have a Z_i which is as-good-as-randomly assigned + excludable

- Assume a distribution for $(Y_i(1), Y_i(0), \nu_i)$ where $D_i = \mathbf{1}[\mu + \pi Z_i > \nu_i]$
- Then we have parametric models for

$$E[Y_i \mid D_i = 1, Z_i = z] = E[Y_i(1) \mid \mu + \pi z > \nu_i] \equiv f_1(z;\theta)$$
$$E[Y_i \mid D_i = 0, Z_i = z] = E[Y_i(0) \mid \mu + \pi z < \nu_i] \equiv f_0(z;\theta)$$

  as well as "first stage" models for $Pr(D_i = 1 \mid Z_i = z) = g(z;\theta)$

- With enough variation in Z_i, the parameter vector θ (and thus ATE) can be identified from these moment restrictions

Key point: the model allows us to extrapolate "local" IV variation to estimate more "policy relevant" parameters

- When Z_i has limited support, the model is doing more "work"
- With full support, we have "identification at infinity" (w/o a model)

Linking Back to LATE

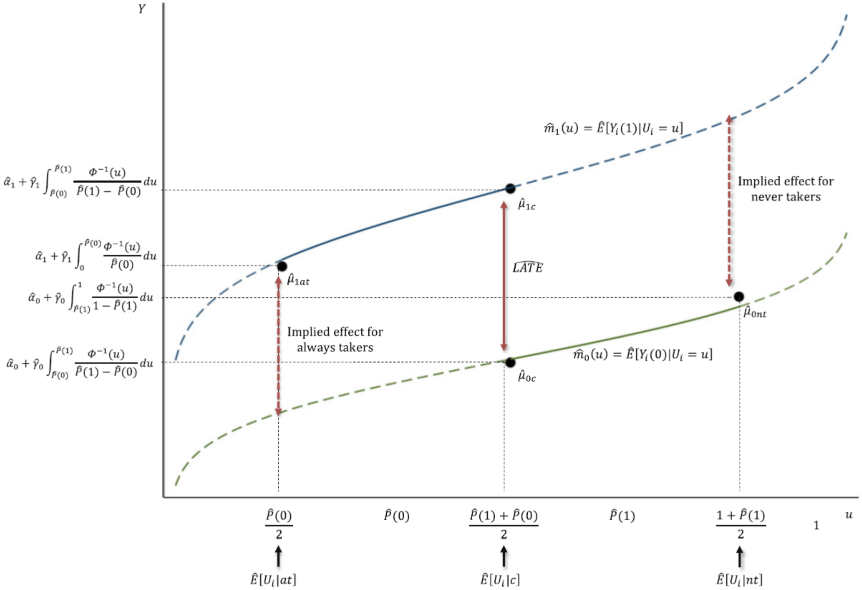
Kline and Walters (2019) formalize this extrapolation logic in the familiar Imbens and Angrist (1994) setup

- Key result: in simple binary Z_i / no controls setup, control function estimates of LATE are numerically identical to linear IV
- "Differences between structural and IV estimates therefore stem in canonical cases entirely from disagreements about the target parameter rather than from functional form assumptions" (p. 678)
- Functional form instead shapes the extrapolation to other causal parameters (similarly to the earlier Probit example)

They conclude with a nice point about validating "structural" models:

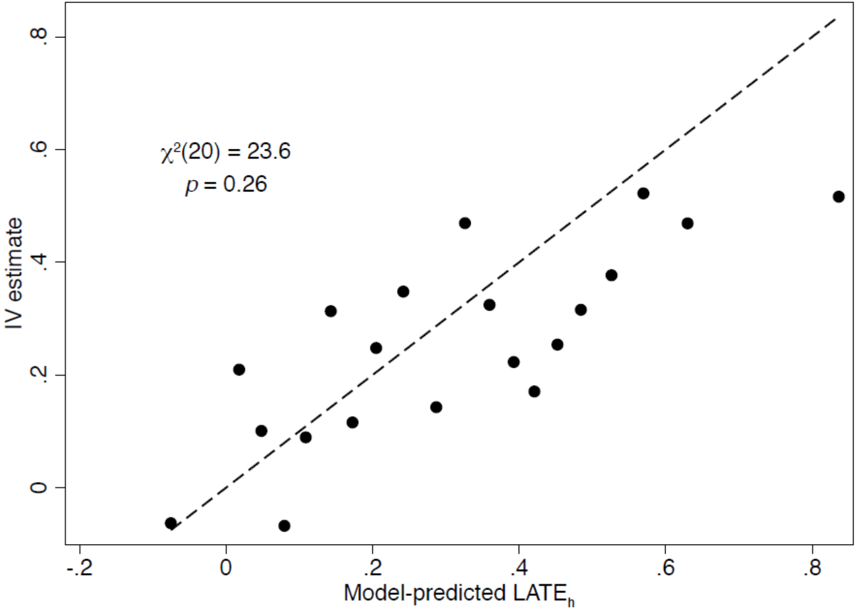
- "Comparing the model-based LATEs implied by structural estimators with unrestricted IV estimates provides a transparent assessment of how conclusions regarding a common set of behavioral parameters are influenced by the choice of estimator" (p. 678)

Heckit Extrapolation of IV Moments



"Heckit" model: $E[Y_i(d)|U_i] = \alpha_d + \gamma_d \Phi^{-1}(U_i)$

Validating Structural Models: Kline and Walters (2016)



Outline

1. Limited Dependent Variables✓
2. Control Functions✓
3. Demand Estimation

Two Goods, Exogenous Prices

Consider a set of markets t, each with population of consumers i deciding between two products $j \in \{1,2\}$ with prices P_{jt} and quality ξ_{jt}

- Firms first enter markets (setting ξ_{jt}), then set prices
- Utility from purchasing good j in market t: $U_{ijt} = \beta P_{jt} + \xi_{jt} + \varepsilon_{ijt}$
- Indicator for purchasing good 1:

$$D_{i1t} = \mathbf{1}[U_{i1t} \geq U_{i2t}] = \mathbf{1}[\beta(\underbrace{P_{1t} - P_{2t}}_{\equiv \bar{P}_t}) + (\underbrace{\xi_{2t} - \xi_{1t}}_{\equiv \bar{\xi}_t}) \geq \varepsilon_{i1t} - \varepsilon_{i2t}]$$

Suppose an RCT randomizes prices, after entry: $\bar{P}_t \perp (\bar{\xi}_t, \varepsilon_t)$. Suppose we further parameterize $\varepsilon_{i1t} - \varepsilon_{i2t} \sim$ Logit. Then market shares are:

$$S_{1t} \equiv E[D_{i1t} \mid \bar{P}_t, \xi_t] = \frac{\exp^{\beta \bar{P}_t + \bar{\xi}_t}}{1 + \exp^{\beta \bar{P}_t + \bar{\xi}_t}},\ S_{2t} \equiv E[D_{i2t} \mid \bar{P}_t, \xi_t] = \frac{1}{1 + \exp^{\beta \bar{P}_t + \bar{\xi}_t}}$$

Regressing $\ln(S_{1t}) - \ln(S_{2t}) = \beta \bar{P}_t + \bar{\xi}_t$ on the randomized $\bar{P}_t$ identifies β

Multiple Goods, Instrumented Prices

Now suppose each market has $J+1$ goods $j = 0, 1, \ldots$. Normalize $\xi_{0t} = 0$

- After entry, a natural experiment randomizes cost shocks C_{jt}. Then firms set prices. Normalize $P_{0t} = 0$. Consumer choices:

$$D_{jt} = \mathbf{1}[\beta(P_{jt} - P_{kt}) + \xi_{jt} - \xi_{kt} \geq \varepsilon_{ijt} - \varepsilon_{ikt}, \forall k]$$

Market shares:

$$S_{jt} = \frac{\exp^{\beta P_{jt} + \xi_{jt}}}{1 + \sum_{k>0} \exp^{\beta P_{kt} + \xi_{kt}}}, \forall j > 0, \quad S_{0t} = \frac{1}{1 + \sum_{k>0} \exp^{\beta P_{kt} + \xi_{kt}}}$$

So again $\ln(S_{jt}) - \ln(S_{0t}) = \beta P_{jt} + \xi_{jt}$

- When $Cov(C_{jt}, \xi_{jt}) = 0$ and $Cov(C_{jt}, P_{jt}) \neq 0$, linear IV identifies β

What Does Structure Buy Us?

We know from Angrist/Graddy/Imbens (2000) that linear IV of (log) quantity on price identifies a "LATE" under monotonicity

- But how useful is this for policy counterfactuals?
- Doesn't generalize easily for multiple "treatments" (products)

With β identified, we can compute welfare (from e.g. new products), conduct merger simulations (given a pricing model), and more

- As before, can think of these as model-based extrapolations: imagine e.g. a binary cost-shock instrument with two products

But... TANSTAAFL. The added structure is not without its cost

- Logit is tractable, but imposes restrictive substitution patterns (IIA)
- We would like to allow for heterogeneity in consumer responsiveness

Berry, Levinsohn, and Pakes ‘95 (Simplified)

BLP is the current “workhorse” model of demand estimation, allowing for flexible substitution patterns while still being (relatively) tractable

Setup: firms enter markets t, pick characteristics (X_{jt}, ξ_{jt}) of products j

- $X_{jt} \in \mathbb{R}^M$ is observed to the econometrician, $\xi_{jt} \in \mathbb{R}$ is unobserved
- Markets have a mass of consumers c,. with characteristics (Q_{tc}, η_{tc})

Given product/consumer characteristics & input costs, firms set prices P_{jt}

- Some versions of BLP parameterize the “supply side”; we’ll skip here
- Market shares S_{jt} are given by consumer maximizing utility:

$$U_{jtc} = \beta P_{jt} + X'_{jt}\gamma + \xi_{jt} + X'_{jt}(\Gamma Q_{tc} + \sigma \odot \eta_{tc}) + \nu_{jtc}$$

  Some versions of BLP incorporate “micro” data on Q_{tc}; here we’ll just assume $Q_{tc} \sim \mathscr{Q}_t$ is known, $\eta_{tc} \sim \mathrm{N}(0, I)$, and $\nu_{jtc} \sim Logit$

Share inversion: $\delta_{jt}(S_t, P_t, X_t; \Gamma, \sigma) = \beta P_{jt} + X'_{jt}\gamma + \xi_{jt}$ for known $\delta_{jt}(\cdot)$

- Note: without random coefficients on X_{jt}, $\delta_{jt}(\cdot) = \ln(S_{jt}) - \ln(S_{0t})$

Price Instruments

Share inversion: $\delta_{jt}(S_t, P_t, X_t; \Gamma, \sigma) = \beta P_{jt} + X'_{jt}\gamma + \xi_{jt}$ for known $\delta_{jt}(\cdot)$

- When $E[\xi_{jt} \mid P, X] = 0$, we can estimate $(\beta, \gamma, \Gamma, \sigma)$ by NLLS
- Inherent price endogeneity: P_{jt} is set strategically, so $Cov(P_{jt}, \xi_{jt}) \neq 0$

BLP note that if we assume $E[\xi_{jt} \mid X] = 0$, we can use instruments $Z_{jt} = f_{jt}(X)$ for known $f_{jt}(\cdot)$; e.g. sums of competitor characteristics

- I.e. using moments $E[f_{jt}(X)(\delta_{jt}(S_t, P_t, X_t; \Gamma, \sigma) - \beta P_{jt} - X'_{jt}\gamma)] = 0$
- See Gandhi and Houde (2020) for discussions of different $f_{jt}(\cdot)$

While common, this restriction can be hard to swallow

- $Cov(\xi_{jt}, X_{jt}) = 0$ is unlikely: what's in ξ_{jt} vs. X_{jt} depends on the data
- $Cov(\xi_{jt}, X_{kt}) = 0$ is also unlikely: strategic entry makes realized characteristics similar within markets

Petrin et al. (2022) solution: add a model for characteristic choice / entry

- Yields other moment conditions $\implies$ other "internal" IVs $f_{jt}(X)$

The Fragility of "Internal" Instruments

Andrews et al. (2022) show that instruments constructed from $E[\xi_{jt} \mid X]$ are inherently sensitive to the assumed model structure

- In contrast to "strongly excluded" instruments (e.g. input cost shocks) C_{jt} that are independent and excludable from potential outcomes (here, potential prices P_{jt} and shares S_{jt})

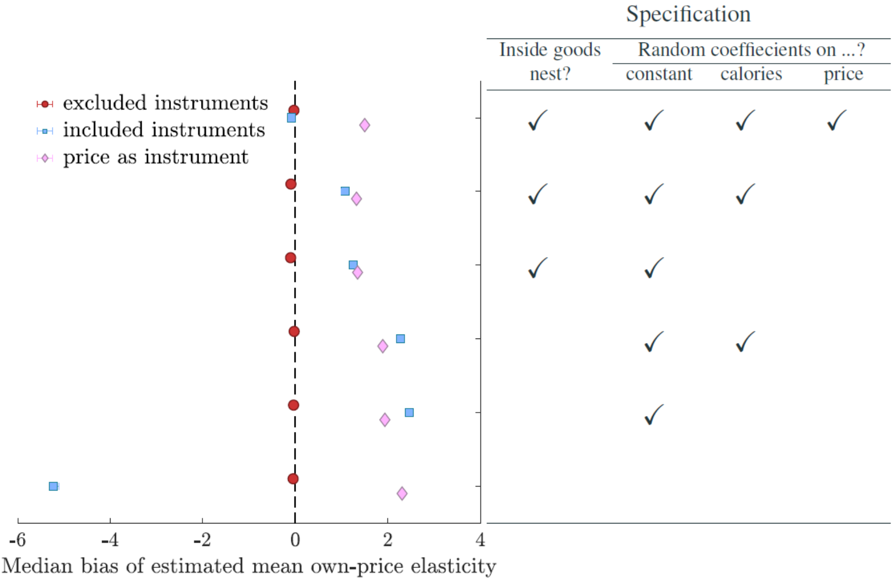
More precisely, they show strong exclusion is sufficient and (essentially) necessary for structural IV estimation to satisfy *sharp-zero consistency*:

- If true effects of P_{jt} on S_{jt} are zero, the IV estimand reflects this
- I.e. that the structural IV identifies some (potentially non-convex) weighted average of true effects

Note: this is a super weak condition! We usually worry when weights are non-convex too (recall Chapter 4)

- But we also know convexity is too high of a bar here: recall linear IV w/multiple treatments

Internal vs. External Instruments: Simulations



Note: simulations based on Miller and Weinberg (2017)

Borusyak and Hull (In Progress)

In ongoing work, Kirill and I are adapting the logic of recentered IV to BLP (and other structural models)

- Key point: even with an excluded C_{jt}, we may need transformations $f_{jt}(C, X)$ to have enough IVs for identification (Berry and Haile '16)
- E.g. sums of cost shocks to close competitors, to identify flexible substitution patterns / "nonlinear parameters"

Moment conditions $E[(f_{jt}(C,X) - \mu_{jt})\xi_{jt}] = 0$, where $\mu_{jt} = E[f_{jt}(C,X) \mid X]$ is given by the "design" of exogenous shocks C_{jt}, similar to before

- E.g. permute exchange rate shocks across countries/periods

Key point: recentered IV can identify structural parameters without any restrictions on the unobservables ξ_{jt}

- While satisfying "sharp-zero consistency" outside of the model

Some Takeaways

There's nothing to fear from nonlinear/structural models, as long as you understand (and are cool with) what the model structure is "doing"

- Likely some extrapolation, which can often be best understood by considering simple / "saturated" versions of the model
- Note linear IV can also be understood as "structural" / extrapolating, just in a way we (often) understand better

To avoid identification "off of functional form," be explicit on what the underlying reduced-form variation is that you're feeding into the model

- Control for / recenter by appropriate controls to isolate it
- Plot model-implied moments vs. reduced-form variation, if you can