Why Panel Data is Indispensable for a Redesigned CE Survey

Jonathan A. Parker,[1] Nicholas S. Souleles[2]
and
Christopher C. Carroll[3]

[1] *Northwestern University and NBER*

[2] *University of Pennsylvania and NBER*

[3] *Johns Hopkins University and NBER*

NBER CRIW Conference
December 2, 2011

Comprehensive Panel *c* Has Enormous Value

- Dramatically expands range, power of feasible analyses
  - Key questions (like response to fiscal stimulus) difficult or impossible to address with cross-section data
  - Price elasticities (and so indexes) better measured with panel
- Error checking across interviews improves data
  - CAPI interviews allow extreme changes from previous levels to be doublechecked in real time; impossible without previous data
- Repeated interviews improve respondent familiarity with process
  - Currently burden is so high that fatigue is more important
  - Preparation and familiarity reduce time and breed accuracy
- Credible panel *c* data in at least one survey allows us to construct estimates of *c* dynamics in other surveys (using, e.g., MPC's out of transitory and permanent income shocks). Without credible panel survey, we have no way to guess about *c* dynamics in *any* survey.

Measuring Expenditures (*a la* Friedman (1957))

- Can't properly measure *y* or *c* over short time span.
  - Consider person who is paid once a month
    - Silly to say that person is "poor" for 29 days and "rich" for 1
    - Friedman: Need ways to measure "permanent" income
    - Friedman: "permanent" *c* is precisely a measure of "permanent" *y*
  - But F notes that there are temporary shocks to spending too
    - Suppose people used to go to local grocery every few days
    - Now much more shopping in occasional trips to "big box" stores
    - Measuring *C* for only two weeks will show greater "inequality" now
    - But that's not *real* consumption inequality
    - It's just like the "poor 29 days, rich one day" kind of income inequality!
    - This might explain, e.g., increased inequality in CE 'diary survey'

Friedman (1957) Implications

- "Panel" spending data needs to be:
  - Comprehensive (not just a few categories)
  - Cover a long enough time span (ideally, two years)
- Not a "panel" in the necessary sense if:
  - it's just *c* measured at two instants separated in time
    - Like, spending on October 1 on successive years
    - Or even spending for a given month in successive years
    - Could be heavily influenced by "did I get to the Sam's Club this month"
  - If it's just current *c* and recalled *c*
    - Recall bias would be significant
    - Anchoring bias to the current level of spending
    - Eliminates benefit of checking outliers from one report to the next

General Framework for Studying Expenditures

Represented by the causal impact of variable $X_{h,t}$ for household h and time t on expenditures $c_{h,t}$, described by the relationship

$$\begin{aligned} c_{h,t} &= \beta_0 + \beta_1 X_{h,t} + \varepsilon_{h,t} \qquad \text{Cross-section} \qquad (1) \\ \varepsilon_{h,t} &= \alpha_h + \tau_t + u_{h,t} \end{aligned}$$

Alternatively, one could compare the change in spending over time

$$\begin{aligned} \Delta c_{h,t} &= \beta_1 \Delta X_{h,t} + v_{h,t} \qquad \text{Panel} \qquad (2) \\ v_{h,t} &= \Delta \tau_t + \Delta u_{h,t} \end{aligned}$$

Notice that the individual effect (α) drops out

Advantage: Price Indexes By *Category* Of Person

One new mandate of CE is to help improve measurement of poverty
Suppose BLS is asked to construct a *price index* for "poor"

- With repeated cross-section alone, have to compare baskets for HH's in the 'poor' income group in consecutive periods
  - Of those low income in t, some would be middle income at $t+1$
  - Of those low income in $t+1$, some would have been middle or high income at t (incomes are particularly volatile for low-income people)
- Most economists would endorse *persistently* low *spending* on *necessities* as a better measure of deprivation (*a la* Friedman's "permanent *c*")

Can Imagine Lots of Similar Examples

- Want a survey that can be used for questions not currently anticipated.
- Suppose the BLS were asked to construct a price index for households with *any* characteristic that varies over time or is measured with error. Like, price index for people with "high medical expenses."
- If only cross-section data are available, price index will inevitably be biased (lumping together, say, people with temporarily high expenses because of an accident, with people with permanently high expenses because of disability).
- Need panel data to measure these things.

Suppose airfares go up

- Proper price index needs to measure subsitution effect
- But what if airlines fiddle with frequent flyer programs to fill seats?
- Will appear to be extremely inelastic: $P \uparrow$ but Q flat
- With only cross-section data, impossible to figure out:
  - Might see big drop in flights that doesn't match airlines' data
  - Unresolvable conflict
- With panel data, might be able to figure it out:
  - Suppose big drop in 'spending' from people who previously traveled a lot?
  - But they have away-from-home hotel spending same time as last year's vacation
  - They probably paid with FF miles
  - Mystery solved!

If $E[\alpha|X] \neq 0$, then cross-sectional estimation is biased and inconsistent

- Example: effect of wealth on purchases when impatient households have lower wealth and, conditional on wealth, purchase more
- Impossible to estimate consistently with cross-sectional data alone
  - In cross-sectional analysis, by including vector of Z_h – persistent household-level characteristics – could estimate consistently if Z_h absorbs *absolutely all* variation in α (and still likely less efficient than panel).
  - Ha!
- Shortly: synthetic panels may be consistent . . . under some conditions

Assume $E[\varepsilon|X] = 0$. Compare cross-sectional estimation of β^{CS} with sample size N and first-difference (FD) estimator on panel data β^{FD} with sample size N.

- Asymptotic statistical uncertainty of β_1 smaller in panel data FD estimator iff $\text{var}\left(\hat{\beta}^{FD}\right) < \text{var}\left(\hat{\beta}^{CS}\right)$ where

$$\text{var}\left(\hat{\beta}^{CS}\right) = \frac{1}{N}\frac{\sigma_\alpha^2 + \sigma_\tau^2 + \sigma_u^2}{\text{var}\left(X_{h,t}\right)}$$

$$\text{var}\left(\hat{\beta}^{FD}\right) = \frac{1}{N}\frac{\sigma_{\Delta\tau}^2 + \sigma_{\Delta u}^2}{\text{var}\left(\Delta X_{h,t}\right)}$$

  - The advantages of panel data are greater the more important household-specific effects (α), the more persistent u, and the less persistent X
- If we assume X, τ, and u are i..i.d. over time, then panel data is more efficient if

$$\hat{\sigma}_\alpha^2 > 0$$

that is, *as long as there are any individual effects*

In CE data (2007 and 2008 data) based on $\beta_1 = 0$ (i.e. only a constant):

Expenditures	Ratio of total Var $(\alpha_h + \tau_t + u_{h,t})$ to FD Var $(\Delta\tau_t + \Delta u_{h,t})$
Food	1.06
Log food	1.78
Nondurable	1.79
Log nondurable	2.87
Total	1.88
Log total	2.49

Thus panel data is on the order of root-2 more accurate (in s.e.'s) than cross-sectional analysis (actual benefit depends on application and past performance is no guarantee of future results!)

$$\begin{matrix}\Delta C_{h,t} \text{ or} \\ \Delta \ln C_{h,t}\end{matrix} = Z_{h,t}\theta + \beta \begin{matrix}\textit{Rebate} \text{ or} \\ \textit{I(Rebate)}_{h,t}\end{matrix} + \varepsilon_{h,t}$$

Spending:	Nondurable	Total	Nondurable	Total	Log Nondurable	Log Total
			Using Panel Data: Dollar change or Log change in spending			
ESP	0.121	0.516			2.09	3.24
	(0.055)	(0.179)			(0.94)	(1.17)
I(ESP)			121.5	494.5		
			(67.2)	(207.2)		
			Using Cross-sectional Data: Level or log spending			
ESP	0.246	0.363			4.54	3.73
	(0.072)	(0.185)			(1.27)	(1.44)
I(ESP)			-94.6	-312.0		
			(84.2)	(206.7)		
Percent Bias	103	-30	-178	-163	118	15

For us: Regressions on the bottom use the same sample in cross-sectional form, so the dep var is level or log consumption and the controls add age squared and are number of kids and num of adults instead of changes. All

Panel data allows estimation of dynamic effects

$$\Delta c_{h,t} = \beta_1 \Delta X_{h,t} + \beta_2 \Delta X_{h,t-1} + \beta_3 \Delta X_{h,t-2} + v_{h,t}$$

But so does cross-sectional data if if households surveyed about past X e.g.

$$c_{h,t} = \beta_1 X_{h,t} + \beta_2 X_{h,t-1} + \beta_3 X_{h,t-2} + \varepsilon_{h,t}$$

But recall and anchoring biases could be significantly worse for cross-sectional data

Economic theory often provides identification in panel data and not in cross-sectional data

Typical optimization conditions for consumption, investment, labor supply, etc. decisions of households imply that only *new information* (and price changes) alter behavior. These conditions imply moments or statistical relationships of the form

$$\begin{aligned} c^*_{h,t+1} &= c^*_{h,t} + \theta \Delta p_{t+1} + u_{h,t} \\ &\text{or} \\ \Delta c^*_{h,t+1} &= \theta \Delta p_{t+1} + u_{h,t} \end{aligned}$$

(for example, Δp_{t+1} might represent the change in the real price of goods between two periods, that is, the real interest rate between these periods).

Without true panel data, evaluation of these conditions or estimation of household preferences from these relationships becomes impossible at least at the household level.

By grouping repeated cross-sections on invariant characteristics, a reseacher can track group averages over time and conduct panel analysis for cohorts as unit of observation

$$\Delta\overline{c_{c,t}} = \beta_0 + \beta_1 \Delta\overline{X_{c,t}} + \overline{v_{c,t}} \quad \text{Panel} \quad (3)$$

$$\overline{v_{c,t}} = \overline{\Delta\tau_t} + \overline{\Delta u_{c,t}} \quad \text{Data} \quad (4)$$

Does not solve shortcomings of lack of panel data.

- Lack of power: depends on how much variation in key dependent variable and error term is collapsed away, on var $\left(\Delta X_{h,t}\right)$ and var $\left(v_{h,t}\right)$ v.s. var $\left(\Delta \overline{X_{c,t}}\right)$ and var $\left(\overline{v_{c,t}}\right)$
- Identification: lose variation in ΔX_c not common to cohort
  - Example: the effect of unemployment on spending;
  - Much more variation in u across individuals than across cohorts
  - Cohort variation is correlated with age which affects spending patterns
  - Eliminates best possible source of variation: the more unrelated to households' characteristics an independent variable is, the less its effects are identified!
  - For some applications var $\left(\overline{\Delta X_{c,t}}\right) \to 0$ with the size of the cohorts: there is no exploitable variation
  - Example: randomized experiments like study of tax rebates (variation across cohorts would be due to differences in eligibility not randomized)

Synthetic Panel Woes (Cont)

- Example: Demand elasticities for price indexes
  - Some prices might vary a lot at the household level but much less at the cohort level (e.g. airline tickets)
- Hard to study populations that change over time: e.g. consumption of stockholders (for estimating, say, “wealth effects”; or homeowners, for estimating spending effects of housing crisis)
  - Less statistical power and require important additional information: Attanasio, Banks and Tanner (2002)
- Only a few examples work, where the variation is aggregate:
  - Effect of change in the minimum wage on spending with variation across US states and time (variation is not lost collapsing across US states): Aaronson, Agarwal and French (2011)
  - Effect of change in after-tax real interest rates across time (variation across time and and across households taxes vary by household characteristic): Attanasio and Weber (1995)

Conclusions

- Redesigned CE survey should focus on those things that it can uniquely do that other surveys cannot.
- Leading example is panel data on spending.
- Can’t even measure price indexes in a credible way if spending data are not credible
- Spending data not credible if they are not measured over time