Models for Time-Series-Cross-Section Data: INTRODUCTION

July 5-8 2011 Workshop at Texas A&M, Department of Political Science

I. **Introductions**…franzese@umich.edu, www.umich.edu/~franzese

II. **Acknowledgements**: these notes & this mini-course draw heavily from materials of similar courses by Greg Wawro, Vera Troeger, & Neal Beck.

III. **Review Syllabus & Logistics**…

A. Tuesday: Intro & Overview; "Fixed" Models of Heterogeneity

1. Intro & Overview

a) Panel & TSCS Data: Definitions, Opportunities, Challenges

b) Background Refreshers [Reviews not covered in lectures; see notes online]

(1) Matrix Algebra, Basic Calc, Prob & Stats

(2) Linear Models - C(N)LRM & G(N)LRM, Estimators - OLS & FGLS

(3) MLE and Logit/Probit models for Binary outcomes

c) Overview of issues & strategies for Panel & TSCS data

2. Modeling Heterogeneity in the Systematic Component
    a) LSDV and Fixed-Effect Models & Estimators (Nuisance)
    b) Interaction Models, Estimation, Evaluation, & Interpretation (Substance)

B. Wednesday: Modeling Heterogeneity in Stochastic Components
    1. Consistent Estimated-Coefficient Variance-Covariance (HAC) Estimation
    2. Random-Effect & Random-Coefficient Models
    3. Testing Random *vs.* Fixed etc., Hybrid Models

C. Thursday: Temporal, Spatial, & Spatiotemporal Dynamics
    1. Models of temporal dynamics & their interpretation
    2. Issues in smaller-*T* samples: Hurwicz/Nickell Bias & panel-IV redresses
    3. Models & Methods for Spatial Association & Spatial Interdependence

D. Friday: 1. Models for Limited & Qualitative Dependent-Variables
    1. Panel/TSCS Models for Limited & Qualitative Dependent-Variables

IV. What is TSCS?

A. Repeated obs…

1. …across *N* units (panels/cross-sections/space): e.g., countries, regions, dyads, firms, parties, groups, individuals, etc.

2. …over *T* time-periods (panels/cross-sections): e.g., decades, years, months…)

3. In particular:

   a) We observe same units, repeatedly, i.e. >1 times.

   b) And, generally, in TSCS *N* not (very) large & *T* not (really) small

B. Data structures & (clumsy, not unified) terminology; a nested scheme (starting most general: my scheme, not standard):

1. *Multilevel/Hierarchical/Nested Data:*

   a) multiple sub-unit observations per unit: y_{ij}

   b) Examples:

      (1) students w/in classrooms;

      (2) survey respondents w/in provinces (&/or states &/or countries);

   c) >2 levels also possible:

      (1) students w/in classrooms w/in schools;

      (2) survey respondents w/in states w/in countries.

2. *Time-Series-Cross-Section (Longitudinal) Data:*

a) Hierarchical where sub-units are time periods: y_{it}

b) Variants: (n.b., these my classification, not standard)

(1) units same each time period (e.g., countries over time, true survey panels…);

(2) or new sample @ period (i.e., a.k.a., *repeated cross-sections*)

3. *Panel (Longitudinal) Data:*

a) Sub-units are time periods; units same each time period.

b) (My distinction, more usually Panel v. TSCS as follows)

4. Most distinguish *Panel* vs. *TSCS* data more by $N \times T$ dimensions:

a) Larger-N, smaller-T = *panel data* & larger-T (smaller-N) *time-series-cross-section data*. But many inconsistencies…

b) Formally, logically, notationally, no diff $N>T$, $T>N$, by how much, etc., but dims crucial to estimation-strategy practical-efficacy since some work well as $T\to\infty$, some as $N\to\infty$ or $NT\to\infty$ or, for some considerations, some particular $f(N,T)\to\infty$.

(1) Many ***panel*** methods designed to redress an *incidental-parameters* problem in N; i.e., the number of parameters growing along with N so consistency lost. Less of an issue in TSCS; parallel there would be incidental parameters in T, but much less attention.

(2) Conversely, with small T, not much can say/learn about temporal dynamics; concern focus instead on distinguish '*sticky* from *stuck*'—i.e., some fixed unit-specific component.

c) May sometimes also care about the units *per se* in TSCS data: states, parties, etc., of intrinsic interest. Care about units *per se* in panel models much less frequently, more usually just random sample of units, no particular interest in individual or firm #1147.

(1) However, resist temptation to confuse "all the data one could possibly get in our one actual TSCS of comparative history" with the population to which we are trying to infer.

(2) That is, note & recall the difference b/w *Descriptive Inference* vs. *Causal Inference*: Questions of empirical fact (history, photographs)—what happened there?—*vs.* questions of empirical support for, manifestation of, theoretical relationships between variables.

5. Alternative, more-standard terminology scheme:

a) ***Multilevel***: subunits not, or not necessarily, time periods;

b) ***Panel*** vs. ***TSCS*** only distinguished by $N \times T$ dimensions;

c) Not common-unit Panel or TSCS becomes ***Repeated Cross-Sections***.

6. Notes; Further Complications:

a) Terminology not consistent across disciplines, or even authors, or even[2] w/i authors!

b) Irregular periodization, not common periods/periodization [headache; won't address]

c) Non-rectangular datasets and missing data [headache; won't address]

d) Combinations of data structures:

(1) Dyadic TSCS $\Rightarrow$ nested units (each dyad involving A within monad A) over time.

(2) Directed-Dyadic TSCS $\Rightarrow$ 3-level nested units (A=sender dyads and A=receiver dyads w/in monad A (i.e., dyads involving A)) over time.

(3) [Somewhat bigger headaches (still manageable); also won't address.]

e) (Possibly nonrandom) unit entry &/or attrition from Panel/TSCS [huge headache…]

C. Example Contexts:

1. *Survey Panels:* NES panels, PSID, etc.
2. *Country-years:* (common in C&IPE, e.g.)
3. *Districts by election-years:* e.g., US Congressional Elects
4. *Legislator-sessions:* votes sequentially
5. Many kinds of *event-history* data
6. *(Directed)Dyad-year* design in IR & parts of IPE
7. *Pseudo-Panels:* Data combining different random samples over time, may group units by "type" across repeated or rolling CS to more closely approximate panel
    a) Compiled ANES surveys, e.g.
    b) Pooled time-series of Democrats, Republicans, and Independents, e.g.

D. **Data Structure & Organization:** Typically, data organized by units, {(unit 1, time 1 to unit 1, time *T*); (unit 2, time 1 to unit 2, time *T*); through to (unit *N* time 1 to unit *N* time *T*)}, as in the following example table, but occasionally convenient to arrange them by time-period instead, {(unit 1, time 1 to unit *N*, time 1); (unit 1, time 2 to unit *N*, time 2); through to (unit 1, time *T* to unit *N*, time *T*)}

country	cc	year	govcons	govconsl	sstran
Australia	1	1961	12.0	11.1	5.9
Australia	1	1962	11.9	12.0	5.7
Australia	1	1994	17.7	18.1	.
Austria	2	1961	12.6	13.0	13.2
Austria	2	1962	12.8	12.6	14.2
Austria	2	1994	18.8	19.0	21.8
Belgium	3	1961	11.9	12.4	11.1
Belgium	3	1962	12.3	11.9	11.5
Belgium	3	1994	15.0	15.0	24.2
Canada	4	1961	15.1	13.4	7.0
Canada	4	1962	14.8	15.1	7.3
Canada	4	1994	20.2	21.5	15.4
Denmark	5	1961	14.4	13.3	7.5
Denmark	5	1962	15.2	14.4	7.7
Denmark	5	1994	25.5	26.3	22.0
Finland	6	1961	11.7	11.9	5.4
Finland	6	1962	12.5	11.7	5.7
Finland	6	1994	22.4	23.3	25.1

V. Opportunities & Challenges of Panel/TSCS Data

A. Substantively interesting opportunities for empirical evaluation:

1. Cross-sectional **&** cross-temporal variation for leverage:

a) Increases number of, & usually (more crucially) useful variation across, observations.

b) Useful for evaluating theories that make predictions in space and time (which is what all theories do…) Enables answer to questions pure CS or pure TS cannot:

(1) *Example 1:* In CS, observe Latino participation-rate in an election is 40%. Could be each given Latino voter is 40% likely to vote, that 40% of Latinos vote every election and 60% never vote, or mix. Can distinguish if observe voters over time.

(2) *Example 2:* Does democracy cause economic development, economic development cause democracy, or do the same conditions that lead polities to democratize or develop also lead them to develop or democratize? Observing countries before & after transitions to democracy—while accounting for (possibly unobservable) factors—can help us evaluate.

(3) *Example 3:* Economies of scale vs. Technological progress (e.g., in Cobb-Douglas / Solow growth model). CS contains info only on former; TS conflates the two; TSCS offers possibility leverage CS for econ. scale and separate tech progress from conflated info in TS.

c) Allows controlling some types of unit (&/or time-period) heterogeneity, even possibly unobserved such het., helping redress omitted-variable bias &/or efficiency

d) Even better, allows (richer) modeling of heterogeneity (substance not nuisance):

e) Even better, allows (richer) modeling of heterogeneity (substance not nuisance):

(1) Is effect of some institution (or drug or policy or…) same in all countries (or persons or firms or…)? If not, why & how does it vary?

(2) Does money-growth have the same effect on output everywhere and "everywhen"? Or does it vary? Latter question more interesting if we ask "and how?" E.g. more nominal contracting => more effect? Pre-Lucas more effect than post?

(3) Does some electoral-institution or election-context or party-systemic feature have same effect on voting behavior of all individuals? How do these effects vary with individual characteristics like education?

f) Allows more explicit, richer modeling of both temporal & spatial dynamics.

2. Examples:

a) *Effects of Bicameralism*: US time-series=virtually no help; US-states cross-section: very little help; US-state TSCS: slightly more help; TSCS of Democracies: much help.

b) *Individual Behavior*: CS: Impossible distinguish impact various aggregate-level factors (e.g., turnout & voting/electoral institutions & condtions); TS: impossible distinguish impact various unit-level factors (e.g., turnout & fixed demographic or SES characteristics of individuals); TSCS=>both become possible.

c) *Institutions*: TS: vary little or none almost by def.; CS: covary w/ all other (relatively) fixed aspects (e.g., "culture"); TSCS: opportunities for leverage greatly enhanced

3. General Principle: (CS: often *little leverage*)(TS: often *little leverage*) => *multiplicatively more than little* leverage; i.e., N & $T = N \times T$, not $N+T$. More to point, potential for useful variation enhanced "multiplicatively" (i.e., greatly).

B. Methodologically interesting challenges for estimation:

1. *Parameter (Model) Heterogeneity*: one of the great opportunities & reasons we want Panel/TSCS data; also, one of its great challenges.

2. *Nonspherical Error Variance-Covariance Structure:*

   a) Conditional variances (skedasticity) unlikely to be constant: *Heteroskedasticity*. This always possible, just especially common & natural to expect it in TSCS data, across/by units (panel *heteroskedasticity*) or over/by time periods (e.g., ARCH).

   b) Observations unlikely to be independent over time: (time-)*serial correlation*.

   c) Observations unlikely to be independent across space: *cross-unit (a.k.a. spatial) (a.k.a. contemporaneous) correlation*.

3. Quasi-philosophical issues of repeated sampling:

   a) *N* fixed, *T* fixed, or just size of sample drawn from (hypothetical) population?

   b) Recall: key question is always population to which you're trying to infer, & relation of sample thereto. Are your aims *descriptive inference* or *causal inference*?

C. Summary of Conclusion re: Strategies w/ these Opportunities & Challenges:

1. *Heterogeneity*:

a) Insofar as you can, model it, usually preferably in the systematic component.

b) Test adequacy of your attempt to model it. To degree you have failed to model it sufficiently, 'control it away' if you can.

c) Insofar as you can, insulate your standard-error estimates from inadequacies in your attempts to model it and/or 'control it away'.

2. *Temporal Dependence*:

a) Insofar as you can, model it, usually preferably in the systematic component.

b) Getting temporal dynamics right is possible, but difficult; may not be essential (almost always is in TSCS dims, less surely so in Panel dims; depends also on aims).

c) Test adequacy of your attempt to model it. Insofar as you can, insulate your standard-error estimates from inadequacies in your attempts to model it.

3. *Spatial Dependence*:

a) Insofar as you can, model it, usually preferably in the systematic component.

b) Getting spatial dynamics right & estimating them properly is possible, though difficult; not always essential (often may be, but depends also somewhat on aims).

c) Test adequacy of your attempt to model it. Insofar as you can, insulate your standard-error estimates from inadequacies in your attempts to model it.

4. *Heteroskedasticity*:

a) Model it (in stochastic component) may add efficiency, but often 2nd-order concern.

b) Test adequacy of your attempt to model it. Insofar as you can, insulate your standard-error estimates from inadequacies in your attempts to model it.

5. Caveats:

a) Econometric procedures to estimate models that properly specify your theoretical and substantive propositions are not always, or even necessarily often, available.

b) Econometric procedures to address simultaneously more than one or two of these potential challenges are not always, or even necessarily often, available.

c) In practice, proposed 'cures' may be worse than 'diseases' because the efficacy of the former and the severity of the latter vary with application specifics.

6. Apply the *Least Distortionary Estimate* principle (Troeger [my paraphrase]): "[Specify, estimate, & t]est your theory in the econometrically best way [you can], but test your theory [i.e., do estimate, interpret, & evaluate it empirically]." [In other words, do best you can: the *Least Distortionary Estimation Strategy*].