Longitudinal Analysis: Introduction to the Course

PS2701-2019

Week 1

Professor Steven Finkel

- Course presents an overview of quantitative methods used for the analysis of longitudinal data, or data collected on multiple units --- individuals, cities, states, countries --- at more than one point in time
- Longitudinal analysis is concerned with the analysis of *change* over time
  - Describing patterns of change for different units
  - Modeling why some units change more than others, and what variables are responsible for those differences
- Focus throughout will be on what has traditionally been the main use of panel data: improving *causal inference* in observational research
- As will see, panel data offers *huge* advantages over cross-sectional data in this regard

Examples

- Why does democracy develop (or break down) in some countries more rapidly than in others?
- What are the consequences of marriage or other life events on partisanship or other political attitudes?
- Do voluntary group memberships promote social and political trust among individuals?
- Do high crimes rates undermine support for democracy?

Kinds of Longitudinal Data

- All longitudinal data can be distinguished from *cross-sectional* data gathered at a single point in time: X_i, Y_i for i=1 to N
- Longitudinal data adds some kind of "t" (for "time of observation") to the mix: X_{it} and Y_{it} for i=1 to N, t=1 to T
- Variations :
  - "**trend data**" or **"repeated cross-sectional data":** X_{it} and Y_{it} at multiple *t* but ***different i*** at multiple *t*
  - "**time series data**": X_{it} and Y_{it} for a ***single i*** at multiple *t*
  - "**panel data**": X_{it} and Y_{it} for ***same i*** at multiple *t*

Kinds of Panel Data

- "**Panel**" Data: Large N, (relatively) small T. This will be our primary focus in the course!
  - Typically ranging from 2-20 time points, though can be larger
  - Continuous Dependent Variable Models
  - Non-Continuous DV Models: Dichotomous, Ordered, Multinominal, Count
  - Panel data can be obtained from surveys, from macro/aggregate data, or from experiments

Kinds of Panel Data (continued)

- "**Time-Series Cross-Section**" Data (TSCS): Small N, (relatively) large T (typically greater than 30).
  - TSCS can be viewed as pooling time-series data over a relatively long time period for a small number of units.
  - Large T allows more complex time-dependent error structures and processes than with small T (panel), and less ability to model cross-sectional variation due to smaller N.
- "**Intensive Longitudinal Data**": diaries, multiple readings per hour, per day, per month of psychic, physical, other states of individuals or dyads (often T>50 or greater, sometimes in the hundreds)
  - As with TSCS, this kind of data is more related to time-series methodology (covered in PS2740) than the shorter time "panel" methods that will occupy most of our attention

Panel Data and Causal Inference

- Begin with cross-sectional regression model:

$$(1) \quad Y_i = \alpha + \beta_k X_{ik} + \varepsilon_i$$

- where X_{ik} are the values of the *k* independent variables for individuals *i* that are thought to "cause" the outcome Y_i, along with a common intercept α and an individual error term ε_i
- Estimate (1) with OLS and interpret the β_k as the "average change in Y for a unit change in X_k"
- What are the problems in terms of causal inference?
  - One problem: with cross-sectional data we interpret the effects in terms of change, but we have no direct measure of *change* in Y, nor *change* in X
  - Panel data obviously overcomes this (mundane) problem, since we have direct measures of change to use in the analysis

- But problems of causal inference in (1) go much deeper, and panel data can help us overcome these problems in many different ways
- Basic Issue: there is a strong likelihood that some or all of the X_{ik} in (1) are *endogenous*, in that they are correlated in one way or another with the error term in ε_i
- This violates the OLS assumption of *exogeneity* of the independent variables, i.e., $E(X\varepsilon)=0$ for all X
- The possibility of **endogeneity bias** in the estimation of causal effects is the fundamental problem in *all* observational research
- Occurs when:
  - Explanatory variables that are related to X have been omitted from the model
  - Y causes X ("reverse causality") in addition (or instead of) X causing Y
  - X contains random measurement errors
- Any of these situations leads to bias in the estimation of the "causal effect" β_k

Omitted Variables

- True Specification: (2) $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$
- Misspecification: (3) $Y_i = \alpha + \beta_1 X_{1i} + \varepsilon_i^*$

  where $\varepsilon_i^* = \varepsilon_i + \beta_2 X_{2i} + \beta_3 X_{3i}$
- X_1 is now related to ε^*, and OLS will yield *biased* (and *inconsistent*) estimates of β_1 to extent that X_1 is related to X_2 and/or X_3.
- Can see this via covariance algebra derivation of OLS coefficients:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

multiply through by X_1:

$$(4)\quad X_1 Y_i = X_1(\beta_0 + \beta_1 X_{1i} + \varepsilon_i)$$

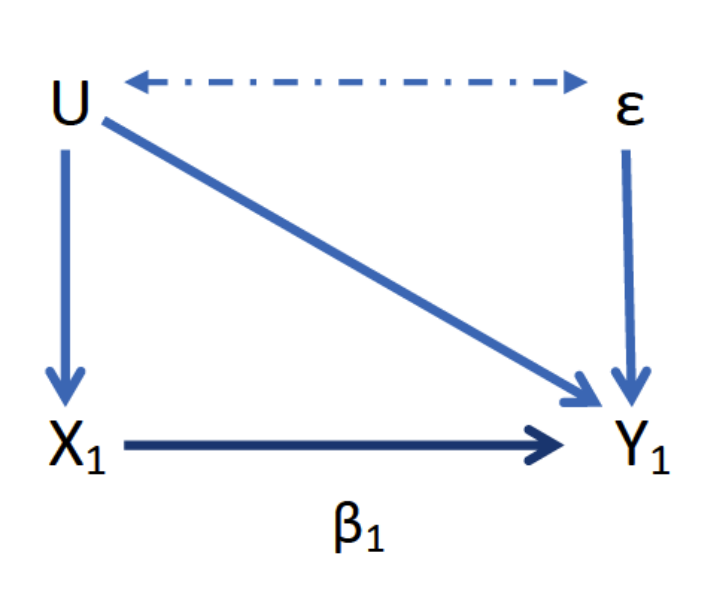
$$X_1 Y_i = X_1\beta_0 + X_1\beta_1 X_1 + X_1\varepsilon_i$$

take expected covariances of each side of the equation:

$$\mathrm{Cov}(X_1 Y_i) = \beta_1 \mathrm{Var}(X_1) + \mathrm{Cov}(X_1 \varepsilon_i)$$

$$\beta_1 = \frac{\mathrm{Cov}(X_1 Y_i) - \mathrm{Cov}(X_1 \varepsilon_i)}{\mathrm{Var}(X_1)} \quad \text{and} \quad \beta_{1OLS} = \frac{\mathrm{Cov}(X_1 Y_i)}{\mathrm{Var}(X_1)}$$

Omitted variable bias in a bivariate causal system



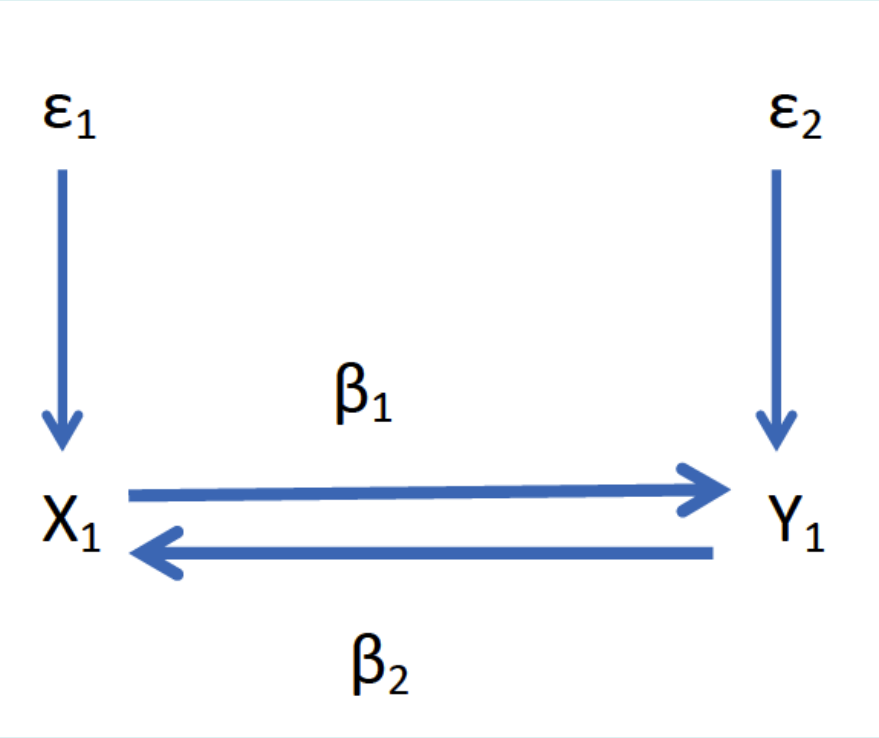
U causes both X_1 and Y_1

U is unobserved, so folded into ε

X is now related to ε and hence "endogenous"

OLS β_1 will give the true effect of X_1 *plus* some correlated effect from U

Reciprocal Effects or Simultaneous Causality



(5)
$$Y_1 = \beta_1 X_1 + \varepsilon_2$$
$$X_1 = \beta_2 Y_1 + \varepsilon_1$$

$$X_1 = \beta_2(\beta_1 X_1 + \varepsilon_2) + \varepsilon_1$$
$$X_1 = \beta_2\beta_1 X_1 + \beta_2\varepsilon_2 + \varepsilon_1$$

So X_1 is related to ε_2, hence endogenous! Same with Y_1 and ε_1 in equation predicting X

Measurement Error in X

Y is related to True X, but we use fallible x_i^* that contains random error v_i

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

(6) $$\text{and } x_i^* = X_i + v_i$$

$$\text{then } Y_i = \beta_0 + \beta_1 (x_i^* - v_i) + \varepsilon_i$$

$$\text{and } Y_i = \beta_0 + \beta_1 x_i^* + (\varepsilon_i - \beta_1 v_i)$$

x_i^* is related to the error term (since v_i and x_i^* are related). So fallible x_i^* is endogenous and we get inconsistent estimates of the causal effect β_1

- Panel data offer multiple ways to overcome endogeneity biases and arrive at more robust causal inference
- Different traditions emphasize different aspects and advantages of panel data in the causal inference process
- All have same goal, though, so should be possible to integrate them!

Approaches to Panel Analysis

- Three major traditions for analyzing panel data, each offering different tools for strengthening causal inference
  - **Econometric Approach** -- rooted in economics and program evaluation, puts greatest emphasis on problem of omitted variable bias, unobserved heterogeneity
  - **Structural Equation Modeling (SEM) Approach** – rooted in sociology, psychology, early political science, puts greatest emphasis on problems of reciprocal causality and measurement error
  - **Multilevel/Hierarchical/Growth Model Approach** – rooted in Biostatistics, Medicine, Education, Developmental Psychology, puts greatest emphasis on temporal change and its determinants
- We will cover each approach, as well as highlight recent efforts at integrating them into a unified analytic framework

Econometric Panel Models

- The econometric tradition emphasizes the use of panel data to overcome the *omitted variable* or "non-spuriousness" problem
- These methods, in addition, embed panel analysis within the more general "counterfactual" approach to causal inference, which has now (or is fast becoming) become the standard framework for conceptualizing and estimating causal effects in the social sciences
- Strategy: Use repeated observations on the same units as a means of estimating the effects of "treatments" (independent variables) that are purged of potential **selection biases** related to *unobserved* or *omitted* variables of different kinds

Cross-sectional model (with unobserved U_i)

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + (U_i + \varepsilon_i)$$

If U is related to any of the included Xs, you are in trouble!

Why? Bias in estimation of the β_j

(And with cross-sectional data there is very little to do about it!)

With panel data: assume (for now) that unobserved U_i is stable

$$Y_{it} = \alpha + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + (U_i + \varepsilon_{it})$$

$$Y_{it} = (\alpha + U_i) + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + \varepsilon_{it}$$

So each unit has its own intercept, determined by U_i !

If U_i is related to the included Xs, potential bias in the β_j!

How can we estimate this model?

- Basic econometric models:
  - "First Difference" (FD) or, in two wave analyses, "difference-in-difference" models to estimate treatment effects
  - "Fixed Effects" (FE): (intuitively), include a "dummy" variable for each unit
  - "Random Effects" (RE): treat U_i as a normally distributed unobserved random variable, folded into a composite error term
  - "Compromise" or "Hybrid" FE&RE Models
- Because of their importance in causal inference and contemporary panel (and social science) literature, we will begin with these models today and next week

Structural Equation Modeling (SEM) Panel Analysis

- Panel data can help overcome endogeneity biases and strengthen causal inference in other ways
  - Can leverage the temporal ordering of variables to see "what causes what"
  - Can use multiple waves of observation to estimate complex models with reciprocal causality and measurement error
  - Panels can sometimes provide "instrumental variables" that can be used to identify some of these effects
- These advantages of panel data are the focus of structural equation modeling (SEM) methods that use panel data to sort out causal direction and correct for measurement error

  Causal Analysis with Panel Data (Finkel, Sage Publications 1995) summarizes these methods; we'll also cover recent work over past several decades that extends the models presented there

- Key features of SEM approach
  - Embeds longitudinal processes within an overall "causal model", a series of equations specifying how variables influence one another over time
  - Specifies a separate equation for each dependent variable at each point in time
  - Usually includes "lagged Y" in each equation (after wave 1), which yields a model that predicts the effects of the X_k on the *changes* in Y, controlling for its previous value
  - Uses information from multiple waves as necessary for identification of reciprocal effects and measurement error parameters

Example: A Two-Wave "Cross-Lagged" Model of Democracy and Social Capital

Paxton(2002), "Social Capital and Democracy:
An Interdependent Relationship", *American Sociological Review*

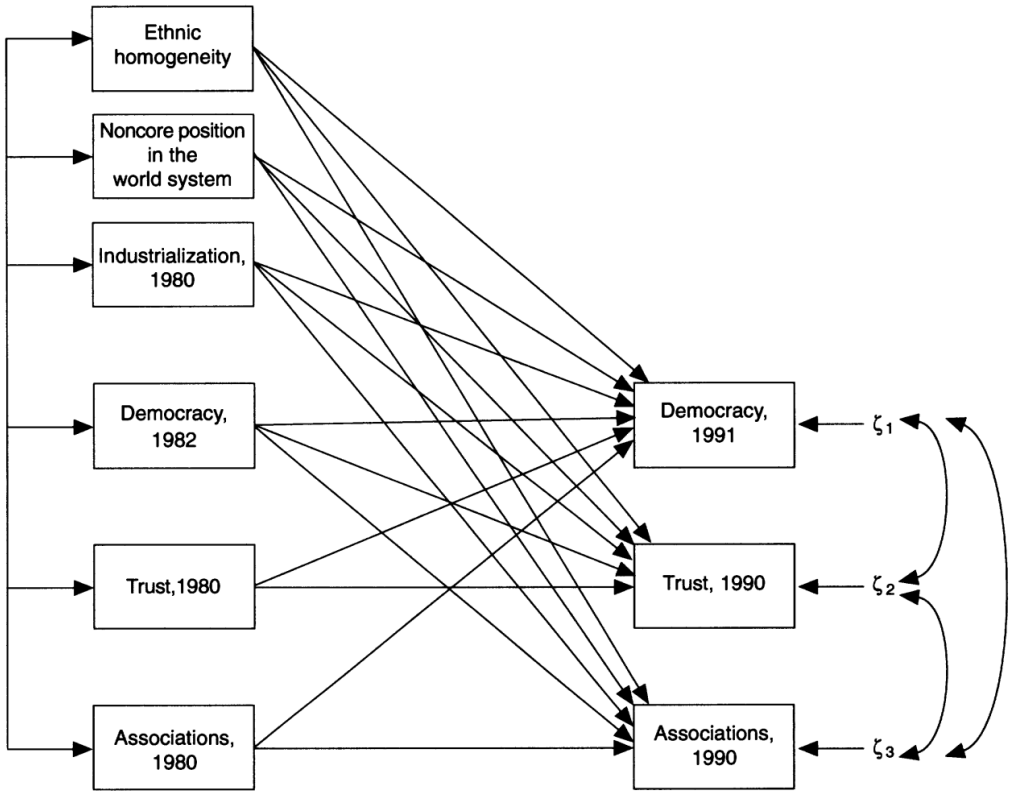
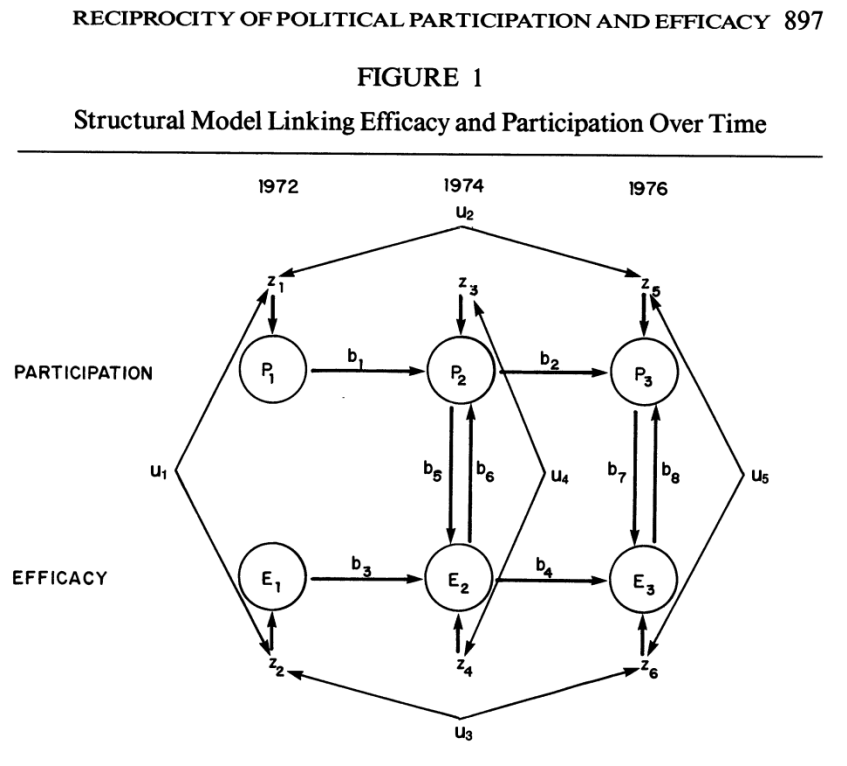


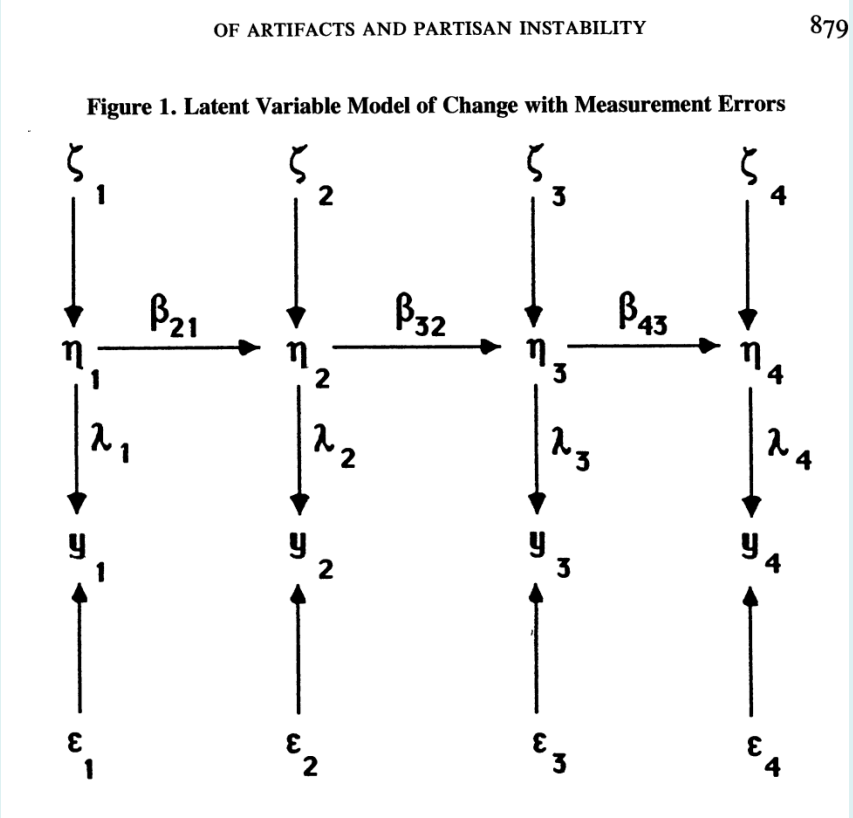
Figure 1. Cross-Lagged Panel Model of Social Capital and Democracy: World Values Survey

Other SEM Panel Analysis Examples

Finkel (1985), *AJPS:* Simultaneous Effects Model Between Political Participation and Political Efficacy



Green and Palmquist (1990), *AJPS:* Single Indicator, Multiple Wave Measurement Error Model of Stability in Party Identification



- Multiple equations representing how X and Y influence one another over time
- Simultaneous estimation of structural effects of entire causal system via STATA SEM module (or via LISREL, EQS, AMOS, MPLUS or, more recently, R's "lavaan" module), less commonly via equation-by-equation analysis
- Variations/Issues:
  - Alternative causal lags: "synchronous" effects and resulting issues in identification and estimation
  - Autocorrelated disturbances
  - Measurement error models
  - All easier with multiple waves, multiple measures

- More advanced work attempts to integrate the econometric and SEM traditions
  - "Dynamic Panel Models" that include both the lagged dependent variable and unit-level heterogeneity (U_i)
  - "Endogeneity" models that attempt to account for potential reciprocal causality and for selection biases due to *unstable*, time-varying factors that may determine selection into a treatment ("endogenous selection")
  - Estimation of these models is possible within ***either*** the econometric or SEM frameworks
  - We'll cover these models at the ends of Units 1 (from the econometric perspective) and Unit 2 (from the SEM perspective)

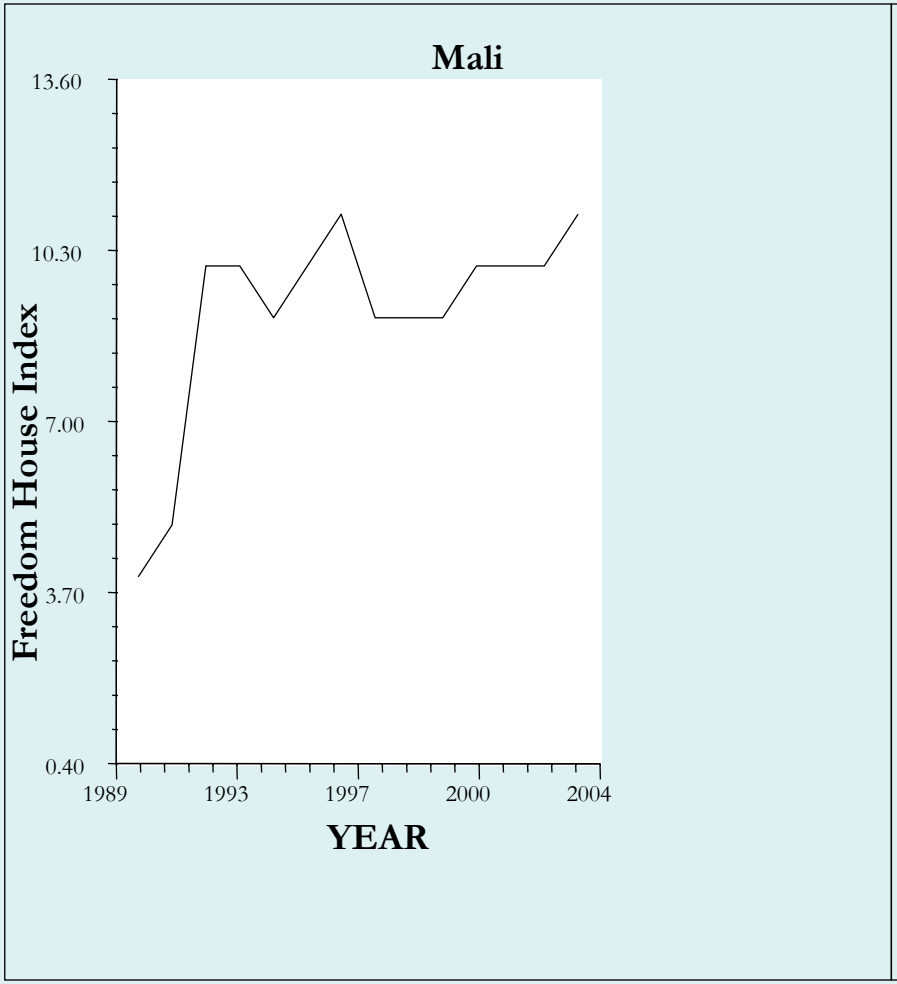
Multilevel and Growth Models

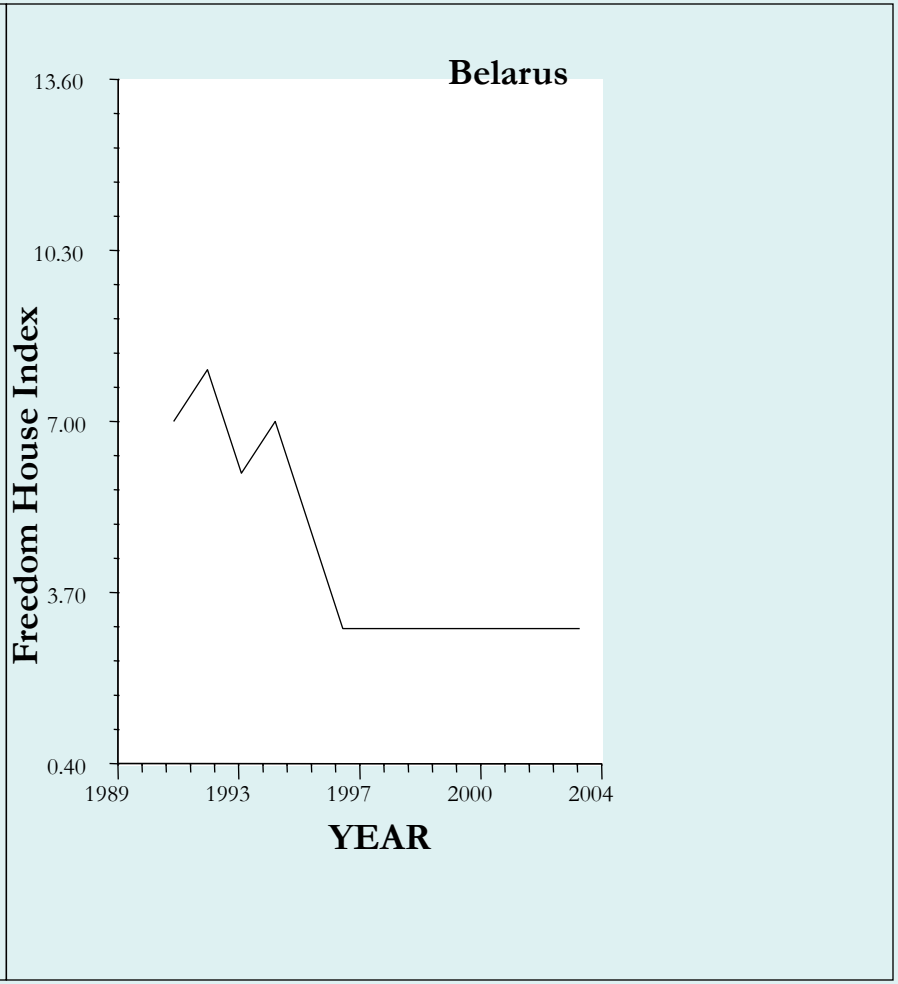
- A final general treatment of panel data has emerged more recently, comes into political science from biostatistics, education, developmental psychology
- Idea of "multi-level" data structures
  - Students "nested" within classes that are "nested" within schools
  - Level 1: *students* Level 2: *classes* Level 3: *schools*
  - Variables exert influence at different levels. Educational outcomes may be the results of:
    - Student-level characteristics (e.g., gender, social class, hours studying)
    - Classroom-level characteristics (e.g. quality of teacher)
    - School-level characteristics (e.g. budget)
  - Observations within each level are "clustered", necessitating special treatment in terms of error structures and intra-level correlations between units

- **Panel data is another kind of multi-level data!**
  - "time" or "wave" of observation is nested within individuals!
  - unit 1 has set of observations at *t1*, *t2*, so does unit 2, 3, etc.
  - Variables at each level may exert influence on outcomes
    - Economic growth for country *i* at time *t* may influence its extent of democracy at time *t* (Level 1)
    - Ethnically polarized countries may have lower levels of democracy at all times (Level 2)
    - The effect of economic growth on democracy may be stronger among countries with less ethnic polarization (Level1-Level 2 interaction)
  - Observations for each country over time are "clustered" and this must be taken into account in the model's error structure

- Major application of multilevel modeling methods to longitudinal analysis: "growth models" that incorporate time as an independent variable
  - Time has effect on some outcome ("growth" at Level 1 for given units)
  - Some units start process at different levels due to unit-level factors at Level 2 (and random error)
  - Some units change or grow at different rates due to unit-level factors at Level 2 (and random error)
  - All level 1 observations within units are clustered, taken into account in error structures
- We'll discuss this approach a bit in a few weeks and in much more depth in Unit 3!

Examples of Democratic Growth Trajectories





Additional Multilevel Panel Models

- Growth models within SEM framework ("latent curve models")
- Models integrating latent growth and lagged DV dynamics
- Random effects and random coefficient models that synthesize aspects of the SEM and multilevel traditions
- Generalized SEM models ("GSEM" in Stata V.13 and later) for multilevel SEM
- Analysis of trend or repeated cross-sectional data via multilevel modeling
  - Multiple higher level units observed over time with *different lower level units* (e.g., cross-national surveys with responses from different individuals in the same country in different years)

Additional Topics: Unit 4

- Models for non-continuous dependent variables
- Longitudinal mediation models and cross-level mediation and interactions
- Models for panel "drop-out" or attrition
- **Unification/integration of the general approaches to panel analysis**

Course Goals

- Understand how to incorporate longitudinal perspective/models in order to answer important research questions and enhance the causal inference process
- Expand your statistical toolkit to include methods that can be used outside of longitudinal analysis as well (Structural Equation Modeling, Multilevel Models)
- Expand ability to absorb and critique longitudinal methods used in published articles and books
- Expand statistical computing skills (STATA SEM and GSEM, XT, and MIXED modules), applications to specialized programs, e.g., MPLUS, AMOS, HLM, LISREL
- More recent applications in R ("plm", "panelr", "lavaan")

Housekeeping Issues

- Requirements
- Class Exercises
- Books
- Stata 15/16/R
- Resources
  - www.stata.com
  - www.paneldata.eu and www.panelwhiz.eu
  - www.electionstudies.org (US National Election Studies)
  - https://www.lissdata.nl (Dutch household panel survey)
  - http://www.diw.de/en/soep/ (German household panel survey)
  - Other Sources: see document in Blackboard, Week 1

Your Professor

- Academic Positions
  - Daniel Wallace Professor of Political Science, University of Pittsburgh, 2005—present. Department Chair, 2011-2018,
  - Professor of Quantitative and Qualitative Methods, Hertie School of Governance, Berlin, Germany, 2005 – 2008
  - University of Virginia, 1984-2005
  - PhD Stony Brook University (year not disclosed!)
- Research Interests
  - Political Behavior, Participation, Democratization
  - Evaluation of Democracy Assistance and Countering Violent Extremism Programs of USAID and International Donors
  - Statistical Methods for Longitudinal and Panel Data
- Short Courses in Panel Analysis
  - University of Bern, Switzerland, January-February 2006
  - Aarhus University, August 2013
  - University of Gothenburg, Sweden, June 2015
  - University of Copenhagen, May 2017
- Home Page: www.pitt.edu/~finkel

Panel Data, Treatment Effects and the Counterfactual Approach to Causality: Introduction to Econometric Panel Models

PS 2701-2019

Longitudinal Analysis

Week 1

Professor Steven Finkel

Panel Analysis, Treatment Effects and the Counterfactual Approach to Causality

- As noted, econometric approaches relate panel analysis to the more general "counterfactual" approach to causal inference
- We use repeated observations on the same units as a means of estimating the effects of "treatments" (independent variables) that are purged of potential **selection biases** related to *unobserved* or *omitted* variables of different kinds ("unobservables" in econometric lingo)
- Intuition: Exploit panel data to use each individual unit as its own control!
- This week, focus on very simple two wave FD/DID "treatment effects" models to illustrate basic principles. Will extend to multi-wave observational panel data with continuous "treatments" and other applications in subsequent sessions. Not all models are developed within the counterfactual framework, but concepts are applicable to all panel models

The "Potential Outcomes" or Counterfactual" Model of Causality

- Developed over the past 50 years and attributed to statisticians Donald Rubin and Paul Holland (and earlier to Jerzy Neyman in the 1920s)
- Virtually all empirical political analysis can in principle be viewed as an attempt to estimate the causal effects of some kind of "treatment" on a particular outcome or set of outcomes
  - Effect of going to college on voting, effect of joining an IO on war, effect of changing electoral laws on number of parties, effects of attending a civic education workshop on political knowledge, etc.
- What do we mean by the "causal effect" of a treatment?
  - Assume a unit can have two "potential" outcomes, depending on whether it is exposed to some treatment *D* or not
    - Y_{1i} is unit *i*'s value of Y if exposed to the treatment (D=1)
    - Y_{0i} is unit *i*'s value of Y if not exposed to the treatment (D=0)
  - So $(Y_{1i} - Y_{0i})$ is the difference in unit *i*'s outcome under the two conditions. It is the difference in the outcome at a given point in time for a unit if it was exposed to D versus the outcome *for that same unit* if it was not exposed to D
  - We call this quantity the *causal effect* of the treatment D

The "Fundamental Problem of Causal Inference"

- Problem: This quantity is *unobservable*!!! We only see one of the two values of Y for a given unit --- Y_{0i} for the control group (D=0), and Y_{1i} for the treatment group (D=1). This is what is known as the **"fundamental problem of causal inference"**!!!
  - We don't know what the control group would have looked like at a given point in time if it had gotten the treatment (Y_{1i}| D=0), and we don't know what the treatment group would have looked like at a given point in time if it had not gotten the treatment (Y_{0i}| D=1). These "**counterfactual**" outcomes are unobserved, so we cannot directly calculate the causal effect of the treatment.

	Treatment Group	Control Group
Treated	**Y_{1i}|D=1 Observed**	***Y_{1i}***|D=0 ***Counterfactual***
Untreated	***Y_{0i}***|D=1 ***Counterfactual***	**Y_{0i}|D=0 Observed**

- Note: panel data doesn't solve *this* problem (though it helps with estimating causal effects under less restrictive assumptions than with cross-sectional data). The *observed change* in Y for the treatment group (D=1), that is, Y at time 2 minus Y at time 1, and the *observed change* from wave 1 to wave 2 in Y for the control group (D=0) are **not** the same as the *change in the potential outcome* $(Y_{1i} - Y_{0i})$ – what we are defining as the "causal effect" – for either group
- Nearly all (modern) empirical social science research is concerned with developing ways of identifying and estimating the unobservable quantity $(Y_{1i} - Y_{0i})$.
- Under what conditions can we use differences between *observed* treatment and control groups as proxies for the true *unobserved* causal effect of interest?

Estimating Treatment Effects in the Potential Outcomes Model: Cross-Sectional Case

- We observe the difference between the treatment group and the control group: $(Y_{1i}|D=1) - (Y_{0i}|D=0)$
- This is ***NOT*** an unbiased estimate of the causal effect except under restrictive assumptions that are not likely to be met in non-experimental research
- If a (potential) causal effect is constant for treatment and control groups, it is easy to show that:

$$E((Y_{1i}|D=1) - (Y_{0i}|D=0)) = E((Y_{1i}|D=1) - (Y_{0i}|D=1)) + E((Y_{0i}|D=1) - (Y_{0i}|D=0))$$

Observed Group Difference = "Treatment Effect on the Treated" + "Baseline Selection Bias"

$$E((Y_{1i}|D=1) - (Y_{0i}|D=0)) = E((Y_{1i}|D=1) - (Y_{0i}|D=1)) + E((Y_{0i}|D=1) - (Y_{0i}|D=0))$$

Observed Difference = "Treatment Effect on the Treated" + "Baseline Selection Bias"

- This means that the observed difference between the treatment and control groups is a function of the (unobserved) causal effect of the treatment on units that get the treatment PLUS the (unobserved) difference in the "no-treatment" outcome between the treatment and control groups.
- The latter term is the difference in what the treatment group ***would have looked like*** in the absence of treatment and what the control group ***did look like*** in the absence of treatment
- Whenever the "selection bias" term is zero, or whenever $E(Y_{0i}|D=1) = E(Y_{0i}|D=0)$, then observed differences between treatment and control groups = the causal effect of the treatment

Randomization Solves the Selection Bias Problem!

- If treatment and control groups are **randomly assigned**, then their baseline potential "non-treatment" outcomes are equalized (as are their potential "treatment outcomes"). The observed control group outcome will be (statistically) the same as what the observed treatment group's outcome *would have looked like* had it not received the treatment.

- In other words, randomization equates the treatment and control groups on ***all*** variables --- *both observed and unobserved* --- that could have produced observed baseline differences between the groups

What about in Observational Research?

- Selection bias will not be a problem whenever *all* relevant X variables that are responsible for the baseline potential "non-treatment" outcome differences are measured and incorporated into the model, i.e., when

$$E(Y_{0i} | D=1, X_i) = E(Y_{0i} | D=0, X_i)$$

- This is called "*selection on the observables*", or "ignorable treatment assignment". It means that we can make the treatment and control groups equal by controlling for X variables that determine D, and then we observe the differences in Y among the groups that are equated on X
- Alternative methods for "controlling" for X:
  - "Matching" individual units on X: match treatment unit 1 with control unit that has the closest characteristics on the X variables that you think determine treatment
  - Multivariate regression: is a kind of matching procedure, on a linear, monotonic function of X covariates (that may or may not provide the best match for a given unit)
  - "Propensity score matching" that match units on many Xs simultaneously: matches treatment unit 1 with the control unit that had the most similar overall probability (propensity) of being treated (but was not treated), based on the set of covariates

- Problem: it is often not reasonable to think that we can control for all Xs that determine treatment status in observational research
- For example, in my work, individuals who select into attending civic education workshops may have different personality attributes, different levels of motivation, different social network characteristics. In IR, countries that enter into international agreements may have different "cultures" than countries that do not. We may or may not be able to measure these things and include them in the model – for that matter, we may not even know what they are.
- This is the problem of "***selection on the unobservables***"!!
- Assume (for now) that they are fixed, stable attributes of individual units. These *unobservables* (call them "U_i") could influence Y, independent of whatever effects X and D may have. If the treatment group and control groups differ on U_i, then comparing treatment and control group --- even after including a host of observed Xs --- will **not** give us the *causal effect of D*

- The problem in regression format:

$$(1) \qquad Y_i = \alpha + \gamma D_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_j X_{ji} + U_i + \varepsilon_i$$

- With potential correlation between D_i (or X_{ji}) and U_i!
- Methods for handling selection on unobservables:
  - Instrumental variables
  - Heckman treatment effects models
  - **Panel data models, e.g. first difference (FD) models, fixed effects (FE), or fixed-random effects (RE-H) hybrid models**

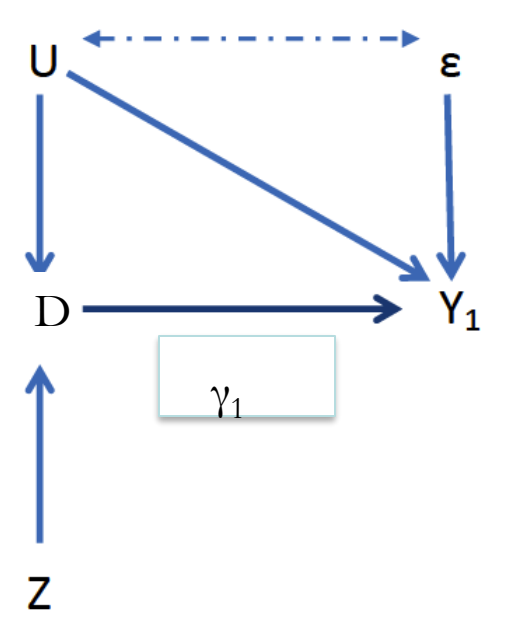
Selection on the Unobservables: Instrumental Variables

$$(1) \quad Y_i = \alpha + \gamma D_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_j X_{ji} + (U_i + \varepsilon_i)$$

- If U_i related to D_i, we have *endogeneity bias* in the estimation of γ, as D_i would be related to the composite error term of the equation
- In terms of the counterfactual framework, we would have baseline selection bias *despite* including all the Xs
  - $E(U_i|D=1)-E(U_i|D=0)\neq 0$, so that $E(Y_{0i}|D=1, X_i)\neq E(Y_{0i}|D=0, X_i)$
- Solution in cross-sectional research: **instrumental variables**
- Find an exogenous Z_i that can proxy for D_i such that:
  - Z_i affects treatment status D_i
  - Z_i is unrelated to any unobserved baseline potential "non-treatment" outcome differences between the treatment and control groups, i.e.

    $E(Y_{0i}|D=1, X_i, Z) = E(Y_{0i}|D=0, X_i, Z_i)$
  - Z_i, unlike D_i, is unrelated to baseline selection bias due to U!
  - Z_i has no direct effect on Y_i; it only affects Y_i indirectly through D_i

Instrumental Variables Analysis, Cross-Sectional Data



To estimate γ_1, we need an instrument Z for D

Conditions that Z *MUST* Fulfill:

1) The "Exclusion Restriction": Z does not cause Y_1 except through D
2) The "Exogeneity Restriction": Z is unrelated to U and ε

- If these assumption hold, then (for an example of a dichotomous Z_i and dichotomous D_i) we can estimate causal effects as:

$$(2) \quad \gamma = \frac{E(Y_i \mid Z = 1) - E(Y_i \mid Z = 0)}{E(D_i \mid Z = 1) - E(D_i \mid Z = 0)}$$

- Or the average outcome difference in Y between units with Z=1 and Z=0 divided by the difference in the proportion of those treated for units with Z=1 and Z=0. This is the so-called "**Wald estimator**" of causal effects.
- Big Problem: where can such instrumental variables be found? Very difficult to find variables that satisfy the assumptions of the IV method!!
- This is why "natural experiments" are increasingly popular. They are "instruments from nature"– naturally occurring exogenous influence on whether a unit receives treatment status that can be viewed "as if" it was randomly assigned, and *sometimes* may also be assumed to have no direct effect on the outcome
- Another strategy: exploit panel data to overcome selection biases!

Selection on the Unobservables: Panel Data

- Panel data offers a wide range of alternative methods for estimating the causal effects of "treatments", taking selection on the unobservables into account
- Basic Strategy: Use longitudinal data to transform the problem from one of possible selection bias due to differential **levels of stable unobservables** for treatment and control groups to one of possible selection bias due to differential **rates of potential "no-treatment" change** over time between the treatment and control groups.
- Estimation models will differ depending on the number of waves of observation and the timing of treatments
  - Two wave quasi-experimental panel designs: "difference in differences" or "first difference" models
  - Multi-wave, time-varying and continuous treatments: first difference, fixed effects, or random/fixed effects hybrid models

The Two Wave Quasi-Experimental Set-Up

	Pre-Treatment	Post-Treatment	Difference
Treatment Group	$Y_{0(D=1)}$	$Y_{1(D=1)}$	$Y_{1(D=1)} - Y_{0(D=1)}$
Control Group	$Y_{0(D=0)}$	$Y_{1(D=0)}$	$Y_{1(D=0)} - Y_{0(D=0)}$
Difference	$\Delta Y_{0(D=1-D=0)}$	$\Delta Y_{1\ (D=1-D=0)}$	$(Y_{1(D=1)} - Y_{0(D=1)}) - (Y_{1(D=0)} - Y_{0(D=0)})$

- This is a common two-wave panel set-up. We estimate the effect of some kind of intervention between time 1 and time 2 that may affect some units (the "treatment group") but not others (the "control group")
- What we are after is "γ" – the "*causal effect*" of the treatment.
- We assign units non-randomly into treatment and control groups, or units select themselves into the treatment and control groups
- We observe outcome *differences* over time for the treatment and control groups, not simply outcome *levels*
- Does this solve the "fundamental problem of causal inference"? Answer: Sort of!!!

	Pre-Treatment	Post-Treatment	Difference
Treatment Group	$Y_{0(D=1)}$	$Y_{0(D=1)} + t_{(D=1)} + \gamma$	$t_{(D=1)} + \gamma$
Control Group	$Y_{0(D=0)}$	$Y_{0(D=0)} + t_{(D=0)}$	$t_{(D=0)}$
Difference	$\Delta Y_{0(D=1-D=0)}$	$\Delta Y_{0\,(D=1-D=0)} + \Delta t_{(D=1-D=0)} + \gamma$	$\Delta t_{(D=1-D=0)} + \gamma$

- What is the difference in Y over time for the two groups?
  - For the control group: a "time" effect that may have changed them ($t_{(D=0)}$)
  - For the treatment group: a "time" effect that may have changed them ($t_{(D=1)}$) , plus the "treatment effect" γ that may have changed them too
- What is the "difference in the differences" (**DID**) between these two groups?
  - The difference in their respective time effects plus the treatment effect on the treated!
- Therefore, analyzing differences (as opposed to *levels*) means:

(1) We have subtracted out any pre-existing observed differences between treatment and control groups!! Any baseline (pre-treatment) selection bias– including influence from "stable unobservables" has been removed **– great news!**

(2) The observed **difference in the differences** will represent the *causal effect of the treatment* whenever the respective time effects are equal, i.e. whenever

$t_{(D=1)} = t_{(D=0)}$

- The same idea can be expressed in "potential outcome" language:

  The observed "difference in differences" will represent the causal effect of the treatment whenever:

$$(3) \quad E(Y^1_{(0)i} - Y^0_{(0)i} | D=1) = E(Y^1_{(0)i} - Y^0_{(0)i} | D=0)$$

  where $Y^1_{(0)i}$ represents the "non-treatment" potential outcome at the post-test (time 1) and $Y^0_{(0)i}$ represents the "non-treatment" potential outcome at the pre-test (time 0). From the previous slide, this corresponds to $t_{(D=1)}=t_{(D=0)}$.

- For the control group, this quantity is **observed**, but for the treatment group, it is **counterfactual**. Since $t_{(D=1)}$ and the causal effect γ happen at the same time, they cannot be disentangled, and we cannot directly test this assumption.

- But ***IF*** we can assume that the change in Y that we *did* observe for the control group is the same as what we *would have observed* for the treatment group had it not received treatment, then the observed difference in a panel or longitudinal set-up between treatment and control groups (the "**DID**") is equal to the causal effect of the treatment!!!

- **This is a great benefit of panel data for causal inference in observational studies!**

- It means we have "only" to assume that whatever unobservables that may differentiate the *levels* of Y for the treatment and control groups don't also influence the *rate of change* in Y over time. This is a weaker assumption and more likely to hold!

Regression Estimation of the Two-Wave QE Model

$$(5) \qquad Y_{it} = \alpha_t + \gamma D_{it} + \beta_1 X_{1it} + \beta_2 X_{2it} + \ldots \beta_j X_{jit} + U_i + \varepsilon_{it}$$

$$\text{Time 1: } Y_{i1} = \alpha_1 + \beta_1 X_{1i1} + \beta_2 X_{2i1} + \ldots \beta_j X_{ji1} + U_i + \varepsilon_{i1}$$

$$\text{Time 2: } Y_{i2} = \alpha_2 + \gamma D_{i2} + \beta_1 X_{1i2} + \beta_2 X_{2i2} + \ldots \beta_j X_{ji2} + U_i + \varepsilon_{i2}$$

- In the QE set up, all D_i at time 1=0
- Since U_i is stable (time-invariant), by subtraction we arrive at the **"First Difference" (FD)** model:

$$(6) \qquad \Delta Y_i = \Delta\alpha + \gamma D_{i2} + \beta_1 \Delta X_{1i} + \beta_2 \Delta X_{2i} + \ldots \beta_k \Delta X_{ki} + \Delta\varepsilon_i$$

- or a regression of change in Y against change in the Xs and the indicator for treatment group status (since D_{i2} is 1 for the treatment group; 0 for control)
- Simple, but powerful model! The possible influence of U_i has been "swept out"/partialled out/conditioned out of consideration

Extension: Time-Varying Effects of Covariates

- Two kinds of covariates: those that are stable over time and those that vary. Let's call stable covariates Z and time-varying covariates X. Then the two-wave FD model looks like:

(7)

$$\text{Time 1: } Y_{i1} = \alpha_1 + \quad + \beta_1 X_{1i1} + \beta_2 Z_i + U_i + \varepsilon_{i1}$$

$$\text{Time 2: } Y_{i2} = \alpha_2 + \gamma D_{i2} + \beta_1 X_{1i2} + \beta_2 Z_i + U_i + \varepsilon_{i2}$$

$$\Delta Y_{i1} = \Delta\alpha + \gamma D_{i2} + \beta_1 \Delta X_{1i} + (\beta_2 Z_i - \beta_2 Z_i) + \Delta\varepsilon_i$$

which, since both Z and the assumed causal effect of Z are constant, reduces to the same FD model from the previous slide

- Implication: stable covariates *with constant effects over time* are eliminated from consideration in the FD (and later, FE or "fixed effects") models. **This is a potential problem with FD if those are among the main variables of theoretical interest**

- But: can easily relax the assumption of constant effects of the covariates, this changes the model only slightly
- For stable covariates Z_i:

(8)

$$\text{Time 1: } Y_{i1} = \alpha_1 + \quad + \beta_j X_{ji1} + \beta_1 Z_i + U_i + \varepsilon_{i1}$$

$$\text{Time 2: } Y_{i2} = \alpha_2 + \gamma D_{i2} + \beta_j X_{ji2} + \beta_2 Z_i + U_i + \varepsilon_{i2}$$

$$\Delta Y_i = \Delta\alpha + \gamma D_{i2} + \beta_j \Delta X_{ji} + (\beta_2 - \beta_1) Z_i + \Delta\varepsilon_i$$

$$\Delta Y_i = \Delta\alpha + \gamma D_{i2} + \beta_j \Delta X_{ji} + \beta_3^* Z_i + \Delta\varepsilon_i$$

where $\beta_3^* = \beta_2 - \beta_1$

- The estimated effect of a time-invariant Z in an FD model is the estimated *difference* in its effects over time
- If Z is a dummy variable, you can also say that β_3 is the additional effect of time for the Z=1 group, over and above the $\Delta\alpha$ which affects all units. If Z is continuous, then β_3 is the additional increment to the time effect for each unit change in Z

- Same thing for non-constant effect of time-varying X covariates:

(9)

$$\text{Time 1: } Y_{i1} = \alpha_1 + \qquad + \beta_1 X_{i1} + (U_i + \varepsilon_{i1})$$

$$\text{Time 2: } Y_{i2} = \alpha_2 + \gamma D_{i2} + \beta_2 X_{i2} + (U_i + \varepsilon_{i2})$$

$$\Delta Y_{i1} = \Delta\alpha + \gamma D_{i2} + \beta_2 \Delta X_i + (\beta_2 - \beta_1) X_{i1} + \Delta\varepsilon_i$$

$$\Delta Y_{i1} = \Delta\alpha + \gamma D_{i2} + \beta_2 \Delta X_i + \Delta\beta X_{i1} + \Delta\varepsilon_i$$

- So the additional coefficient on wave 1 X represents the *difference* between the effects of X on Y at wave 1 and wave 2

Alternative Model for QE Estimation

- Here is another set-up for the two-wave QE longitudinal model:

$$(10) \quad Y_{it} = \alpha + \beta_1 D_i + \beta_2 Time_t + \beta_3 D_i * Time_t + \varepsilon_{it}$$

  - It says that Y at a given point in time is equal to: a common intercept, an effect (β_1) of whether the unit is in the treatment group or not, an effect (β_2) of a given time period on all units, an interaction effect (β_3) of time with treatment group status, and an idiosyncratic error term (ε_{it})

- For the two groups at each time point, the equation reduces to:

$$(11) \quad \text{Control, Time 0:} \quad Y_{i0} = \alpha + \varepsilon_{i0}$$

$$\text{Control, Time 1:} \quad Y_{i1} = \alpha + \beta_2 + \varepsilon_{i1}$$

$$\text{Treatment, Time 0:} \quad Y_{i0} = \alpha + \beta_1 + \varepsilon_{i0}$$

$$\text{Treatment, Time 1:} \quad Y_{i1} = \alpha + \beta_1 + \beta_2 + \beta_3 + \varepsilon_{i1}$$

- Taking differences means that the control group change over time is β_2, the treatment group change over time is $\beta_2+\beta_3$, and the "difference in difference" in the two groups is β_3, which represents the causal effect of the treatment.
- So β_3 in equations 10-11 is the same as γ in equations 5-9!
- This means that you can recover the "difference in differences" effect *either* through a true difference model, OR through a regression model with treatment group status, time, and an interaction effect of treatment group status and time. Same result!
- Interestingly, this means that you do not ***need*** panel data on the exact same units to estimate the treatment effect in QE kinds of analyses: you only need randomly selected treatment and control units at two points in time, but the units could be different units within the treatment and control group populations
- Panel data provides additional information on changes at the individual level among units with **direct experience** with treatments that cannot be obtained through other means, however

Examples:

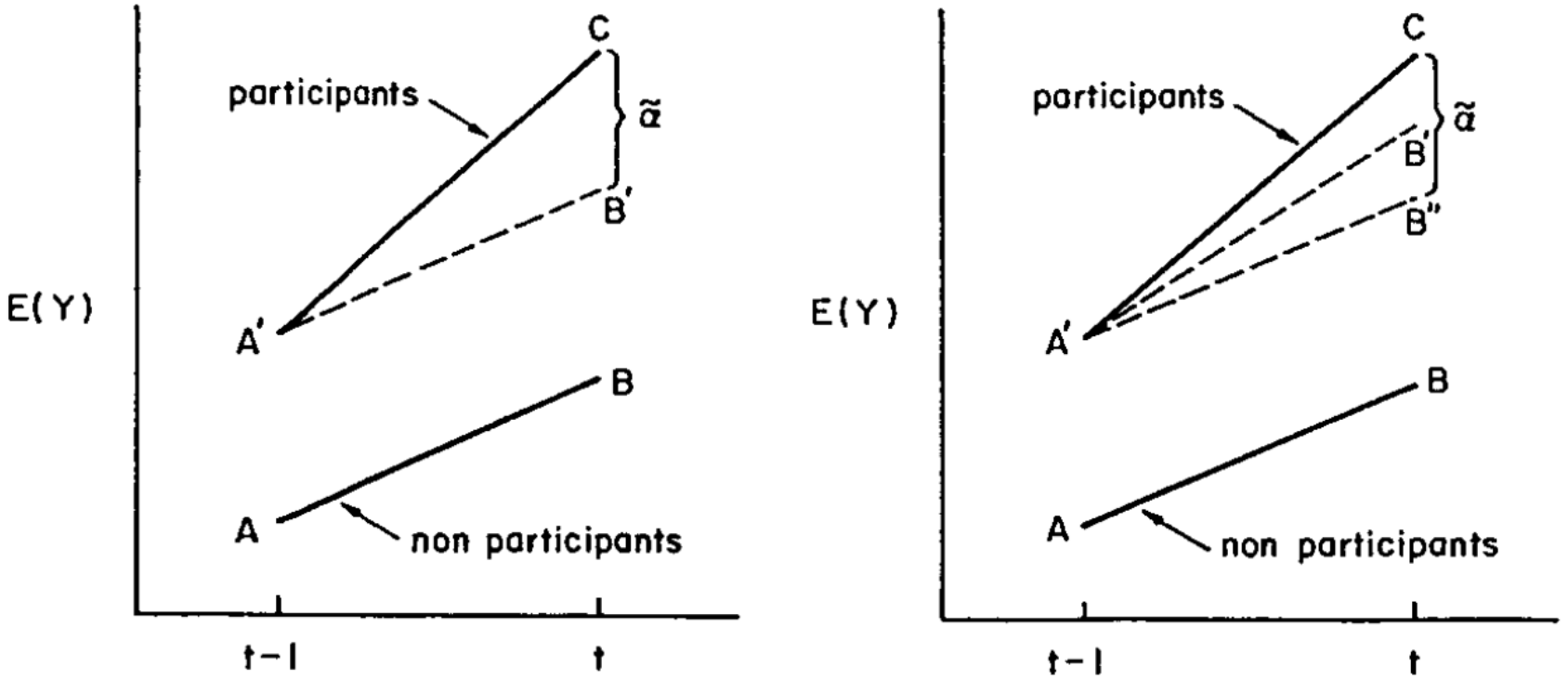
- Woolridge discusses estimating of causal effects of a new garbage incinerator on housing prices in Massachussetts via DID: look at housing prices in the region near the incinerator *before* and *after* the incinerator was built, compare to housing prices outside the region before and after the incinerator was built.
- Angrist and Pischke discuss the effects of minimum wage laws on employment via DID: look at employment before and after the minimum wage increase in a sample of restaurants in a state (NJ) that imposed a minimum wage increase versus a sample of restaurants at the same time points in a state (PA) that did not.
- In both cases the same units were not observed, but the "treatment" was carefully defined, as were "treatment" and "control" samples
- Estimating the model in this fashion requires "long" format data (see next week); the standard model can be estimated with either "long" or "wide" format data

Testing FD Assumptions

- FD model will be valid when the change over time observed for the control group ---($\Delta\alpha$ in equation 6)-- is the same as what the change over time *would have been for the treatment group had it not been treated*. In that case γ in equation 6 provides the causal effect we want. In counterfactual language, are counterfactual rates of change in the "no-treatment" potential outcomes likely to be the same for the treatment and control groups, i.e.

  is $E(Y^1_{(0)i}-Y^0_{(0)i}|D=1)= E(Y^1_{(0)i}-Y^0_{(0)i}|D=0)$?

- It may be that whatever unobservables distinguish the "pre-treatment" values of the treatment and control groups also would lead to *differential changes over time*. This used to be called a "selection-maturation effect" that could bias inferences, as the treatment group was already "maturing" at a faster (or slower) rate and might select themselves into treatment as well. If so, we won't get the causal effect we want.



- Panel (a) is fine. In panel (b), though, the treatment group would have changed more than the control group, even in the absence of treatment, so FD overestimates the treatment effect as (C-B") instead of (C-B')
- With two wave data, impossible to do anything about this, but with at least three-wave data, can begin to make headway

- Possible Strategy: model the underlying time trajectory for both treatment and control groups separately and include this trajectory as a further control for the time-specific treatment effects.

$$(12)\ \mathrm{Y}_{it} = \alpha + \gamma D^{*}_{it} + \beta_1 X_{it} + \beta_2 D_i + \beta_3 \mathrm{T} + \beta_4 (D_i \mathrm{T}) + \varepsilon_{it}$$

- where D is an indicator for whether the unit is in the treatment or control group, T is the time trend, and X are the observed covariates.
- Here the crucial effects are β_3, the effect of time for the control group; β_4, the effect of time for the treatment group; and γ, the effect of being treated at a given point in time, over and above the "pre-existing" time trends for the two groups.
- This is a *very* conservative test – "controlling" for D_i and $D_i T$ assumes that none of their effects is a "true" treatment effect
- Need many waves and units with non-constant treatments to make this a robust test

TABLE 1 Sample Design for Longitudinal NCEP Evaluation

	Wave 1 February–April 2002 (A)	Wave 2 October–November 2002: Follow-up from Wave 1 (B)	Wave 3 April–June 2003: Follow-up from Wave 1 Only (C)	Wave 3 April–June 2003: Follow-up from Waves 1 and 2 (D)	Total Respondents: All Two- or Three-Wave (B+C)	Total Respondents: Two-Wave Only (B+C-D)
Initial Workshop Attendees	1308	901	261	210	1162	952
Initial Nonattendees	1303	886	253	191	1139	948
Total	*2601*	*1787*	*514*	*401*	*2301*	*1900*

TABLE 2 Two-Wave Fixed Effect and Three-Wave Differential Trend Models: Civic Education's Effect on Democratic Orientations

	Political Knowledge		Participation		Tolerance		National vs. Tribal Identification	
	Two-Wave Fixed Effects (a)	Three-Wave Differential Trends (b)	Two-Wave Fixed Effects (a)	Three-Wave Differential Trends (b)	Two-Wave Fixed Effects (a)	Three-Wave Differential Trends (b)	Two-Wave Fixed Effects (a)	Three-Wave Differential Trends (b)
Civic education exposure	0.120**	0.117**	0.106**	0.100**	0.129**	.033#	0.071**	.082**
	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.01)
Media consumption	0.520**	1.342**	0.447**	0.561**	0.231*	.455**	0.090	0.161**
	(0.08)	(0.06)	(0.15)	(0.11)	(0.09)	(0.07)	(0.06)	(0.05)
Political interest	0.048	0.088	0.505**	0.627**	0.036	−0.061	−0.067	−0.092*
	(0.08)	(0.06)	(0.14)	(0.10)	(0.09)	(0.07)	(0.06)	(0.04)
Group memberships	0.441**	0.591**	2.972**	3.235**	0.103	0.054	−0.181*	−0.325**
	(0.11)	(0.07)	(0.19)	(0.12)	(0.11)	(0.08)	(0.08)	(0.05)
General political discussion	.089**	.209**	0.181**	0.263**	0.027	.063**	.054*	0.058
	(.028)	(0.02)	(0.04)	(0.03)	(0.03)	(0.02)	(0.02)	(0.01)
November reinterview	0.318**		−0.690**		−0.232**		0.122**	
	(0.04)		(0.07)		(0.04)		(0.03)	
March-June reinterview	0.013		−0.528**		−0.230**		0.120**	
	(0.06)		(0.10)		(0.06)		(0.04)	
Treatment group		−0.045		0.143*		−0.035		0.050#
		(0.04)		(0.06)		(0.04)		(0.26)
Time trend		0.024		−0.305**		−0.122**		.079**
		(0.03)		(0.05)		(0.03)		(0.02)
Trend × treatment group		0.092**		−0.039		0.086*		0.017
		(0.03)		(0.06)		(0.04)		(0.03)
Constant	1.582**	0.916	0.808**	0.213*	1.730**	1.646**	0.174**	0.175**
	(0.08)	(0.06)	(0.14)	(0.09)	(0.09)	(0.06)	(0.06)	(0.04)
No. of observations	4593	4993	4593	4993	4586	4983	4583	4983
R-squared	0.181	0.214	0.166	0.222	0.021	0.018	0.069	0.056

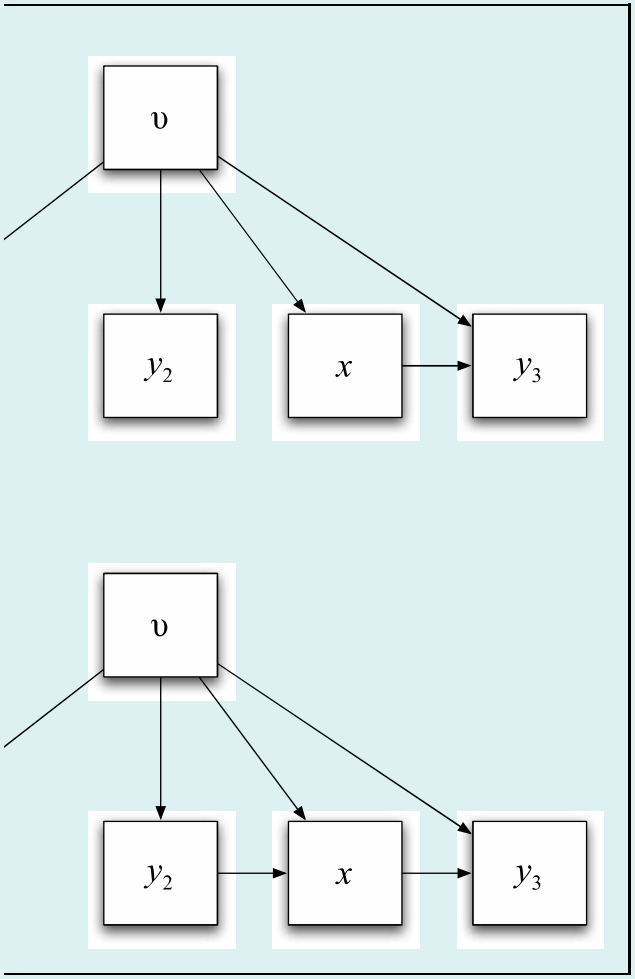
Note: Robust standard errors in parentheses are clustered on 2,301 individuals in all models. Coefficients are significant at #$p < .10$; *$p < .05$; **$p < .01$. *R*-squared within is presented for two-wave fixed-effect models.

Example of FD/FE Panel Models:

Finkel and Smith, Civic Education, Political Discussion, and the Social Transmission of Democratic Knowledge and Values in a New Democracy: Kenya 2002, *American Journal of Political Science* (2011)

- **Last assumption: Is the reason that the treatment and control groups may differ at the pre-test, controlling for X, due to *stable* unobservables?** If so, you can apply the models we have been discussing and subtract them out of consideration. If not, you can't. For example, people may select themselves into employment programs based on a transient dip in earnings at time t-1, not because of "U" (this is the famous "Ashenfelter's Dip" in economics).
- This is also called **"endogenous selection"** into treatment. We need to estimate alternative endogeneity models, some that may include the lagged dependent variable as an additional control. These models require additional waves of observation as well. We will consider these models in later sessions.
- More generally, this means that one needs to think very carefully about the selection process whereby units receive treatment. The simple FD/QE model here assumes that U->D, U->Y_{t-1}, and U->$Y_{t,}$ but there are other possibilities with important analytic implications

Alternative Selection into Treatment Mechanisms



Summary

- We can control for selection biases due to stable unobservables via the FD model, one of the "within" estimators in panel analysis that uses the unit as its own control
- With two-wave data, the model is easily applied to QE type data where control and treatment groups are observed pre- and post-treatment
- Stable *observed* variables with constant effects wash out of the FD model, which is a drawback. The model is flexible enough, however, to allow for *non-constant effects* of both stable and unstable observed variables over time
- FD (and all "within" estimators) assume a common counterfactual time trend in order to identify causal effects; we assume that units that receive the treatment in a given time period, had they not received treatment, would have changed by the same amount in that time period as control units (which *in fact* did not receive treatment)
- This assumption, and the assumption of selection due to *stable* unobservables, can be tested, and possibly relaxed in more complex models with multi-wave data

Sønderskov, Dinesen, Finkel and Hansen, "Crime Victimization Increases Turnout: Evidence from Individual-Level Administrative Panel Data" *APSA* presentation 2018