PS2030
Political Research and Analysis
Unit 4: Models for Causal Inference
3. Selection on Unobservables: Instrumental Variables, Difference-in-Differences
4. Fixed Effects and Panel Models

Spring 2025, Weeks 12-13

WW Posvar Hall 3600

Professor Steven Finkel

Selection on Unobservables

- Assumptions we have made to this point about treatment assignment: it is "ignorable", conditioned on observable X covariates.
- Big Problem: it is often not reasonable to think that we can control for all Xs that determine treatment status in observational research
- As we have discussed over the course of the semester in regards to civic education, e.g., individuals who select into attending civic education workshops may have different personality attributes, different levels of motivation, different social network characteristics. 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***", or "*non-ignorable treatment assignment*". Propensity score matching and other matching estimators can't account for this (except to the extent that the unobservables may be correlated with observed Xs)

- If these *unobservables* (call them "U_i") influence Y, independent of whatever effects X and D may have, and 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 in propensity score or other matching analyses --- will **not** give us the *causal effect of D* since baseline selection bias will still exist
- How to take unobservables into account? One indirect way is through "sensitivity analysis", for example, Rosenbaum's method implemented in STATA with the add-on called "**rbounds**". These methods simulate how large the unobservable variable(s)" effect on the treatment would have to be in order to render the estimate of the causal effect of the treatment statistically insignificant. We won't discuss this in more detail, but see Finkel, Horowitz and Rojo-Mendoza (*JOP* 2012) for an application.

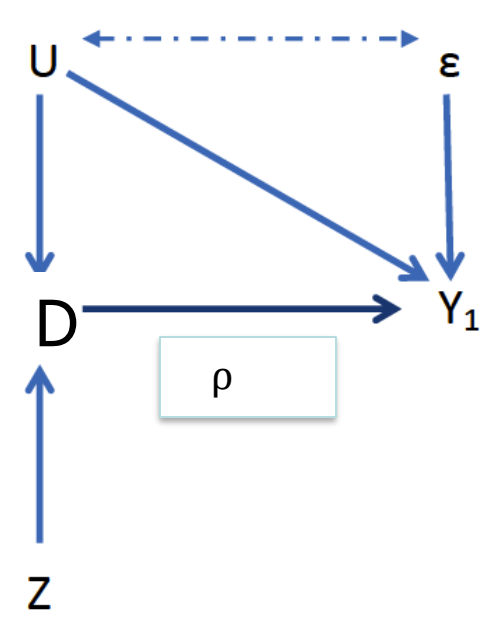
- More direct methods for handling selection on unobservables:
  - Instrumental variables
  - Panel or longitudinal difference-in-differences (DID) models
  - Heckman selection models
  - Regression discontinuity designs
- Recall that week 8 was spent on instrumental variables, so we already have a solid foundation on that topic, even if we have not directly integrated that discussion into the Rubin-Holland potential outcomes causal inference framework
- PS2701 Longitudinal Analysis covers the DID and many other panel models in more detail; some of the other models from this unit (and other more advanced applications) are covered in PS2702 Causal Inference

Selection on the Unobservables: Cross-Sectional Case

$$(1)\quad Y_i = a + r D_i + b_1 X_{1i} + b_2 X_{2i} + \ldots b_k X_{ki} + (U_i + e_i)$$

- If U_i related to D_i, we have *endogeneity bias* in the estimation of ρ, 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 $E(Y_{0i}|D=1)\neq E(Y_{0i}|D=0)$
- 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.
  - 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 ρ, 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:

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

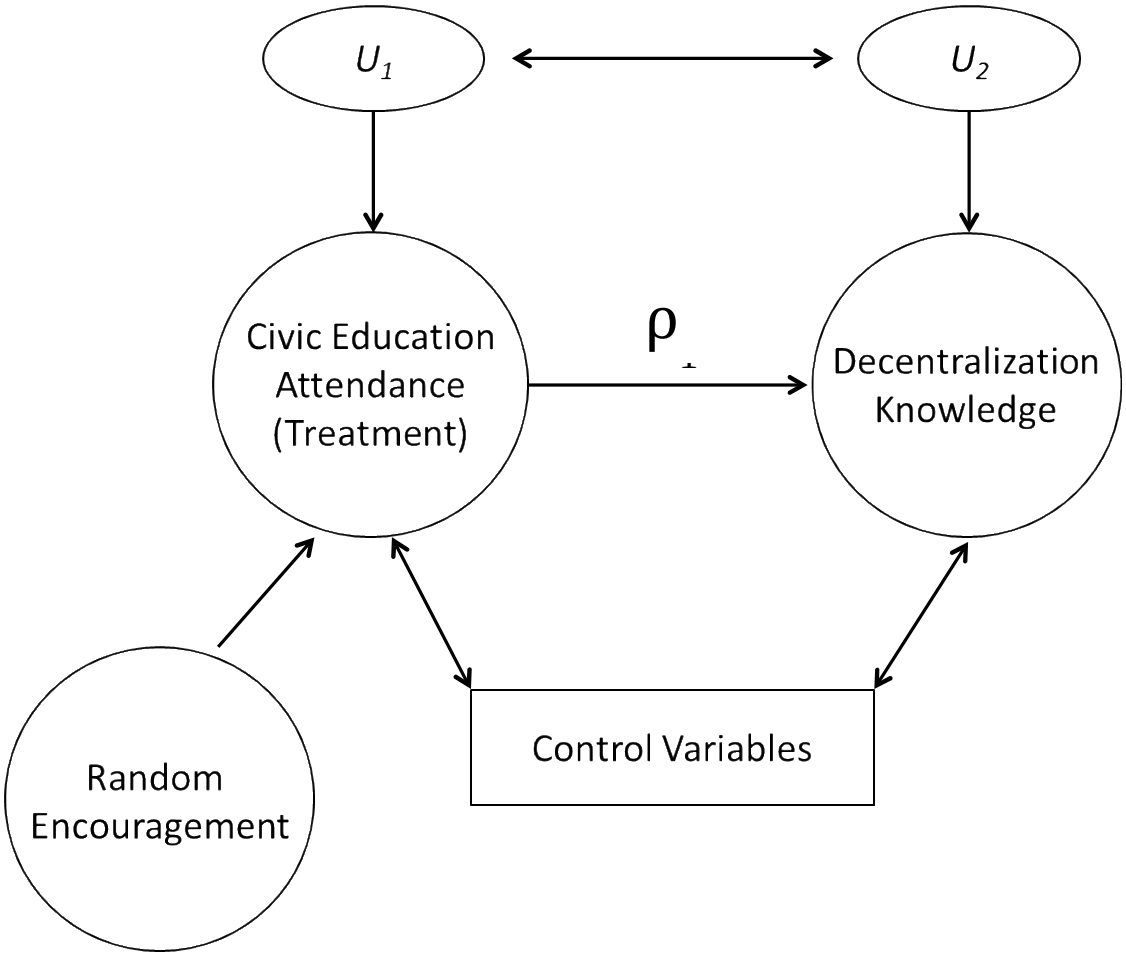
- Or the average outcome difference in Y between units with Z=1 and Z=0 divided by the difference in the proportion of those treated for units with Z=1 and Z=0. This is the so-called "**Wald estimator**" of causal effects.
- Big Problem, as we discussed in week 8: where can such instrumental variables 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

IV and "LATE"

- IVs represent an exogenous source of variation in the treatment that is unrelated to the outcome aside from its indirect role in changing the treatment. Advanced work (following Angrist and Imbens 1994) formalizes this idea in the context of *heterogeneous treatment effects*, i.e., different effects on the treatment for different kinds of individuals.
- Imagine that the world consists of several kinds of individuals/units, each with different treatment effects:
  - Those who take up the treatment *no matter what*, i.e., whether or not Z is introduced
  - Those who take up the treatment *because of* Z being introduced or because they were "pushed" into the treatment from Z
  - Those who would not take up the treatment *no matter what*, i.e., whether or not Z is present
  - Those who *refuse* to take up the treatment if Z is introduced, and *take up* the treatment is Z is not introduced
- Angrist and Imbens label these groups as follows:
  - "Always Takers"; "Compliers"; "Never Takers"; "Defiers"
- It is then shown that IVs identify the treatment effect *only for the Compliers*, those who took up the treatment because of Z! So IVs in the counterfactual framework are now said to identify the "Local Average Treatment Effect" (LATE) or the "Complier Average Treatment Effect" (CACE). **Very important for interpretation!!!**

Example of LATE: Randomization as an Instrument

- Problem in civic education evaluation research: self-selection into the treatment, such that baseline selection bias is unlikely to be ruled out through inclusion of observed covariates.
- Possible solution: randomly "encourage" some individuals but not others in the treatment area to attend the event.
- The "encouragement" is used as an instrumental variable or proxy for actual attendance to estimate the causal effect of exposure to the event
- What is the LATE? It is the effect of the treatment among those individuals who were "pushed" into attending the civic education event *because* of the encouragement. We still cannot identify the treatment effect among those who would have attended regardless of the encouragement (the "Always Takers") or those who would not attend regardless of the encouragement (the "Never Takers").


U_1
U_2
Civic Education Attendance (Treatment)
ρ
Decentralization Knowledge
Random Encouragement
Control Variables

Estimating Causal Effects in the Encouragement Design

- If we examine the raw differences in Y among the encouraged and non-encouraged groups (controlling for X), we get the "reduced form" effect of Z on Y. This is called an ***"Intent to Treat" (ITT)*** effect and is common in the experimental literature. It measures the effect of the randomization on the outcome, whether or not the units "took up" the actual treatment or "complied" with the randomization. Also used in medical studies and many field experiments in political science and economics when there is imperfect compliance with a randomized treatment.
- The ITT is a very good substantive measure of the overall effects of a program when the program itself is randomized, since it is saying, here are the differences on Y between areas/people who were exposed to the treatment, whether or not they *actually* experienced or took up the treatment.
- We can also use encouragement as an IV "proxy" for actual attendance in an instrumental variables analysis. If the encouragement works, it is related to treatment but *unrelated* to every other variable (U or X, stable or unstable) that may be related to baseline differences between the groups.
- But according to LATE, important to remember that the treatment effect is identified *only* for the subgroup of individuals who were moved to attend/not-attend because of the presence or absence of the encouragement

Instrumental Variables (2SLS) Analysis

$3(a)$: $Y_i = \alpha + \rho D_i + U_i + \epsilon_i$ and $\mathrm{E}(D_i U_i \neq 0)$

- Stage 1: Predict D_i from Z (encouragement)

  $3(b)$: $D_i = \tau + \beta Z_i + \upsilon_i$

  with Z_i as instrument and $\mathrm{E}(Z_i U_i = 0)$

  $3(c)$: $\widehat{D_i} = \tau + \beta Z_i$

- Stage 2: Enter predicted D into the 3(a) equation

  $3(d)$: $Y_i = \alpha + \rho \widehat{D_i} + U_i + \epsilon_i$

- **Important: all assumptions and tests for 2SLS from week 8 are also relevant here (Shea, Sargan, etc.)**

Results: Finkel and Lim, "The supply and demand model of civic education: evidence from a field experiment in the Democratic Republic of Congo", *Democratization (2021)*

Table 2. Effects of voice exposure on perceived democratic supply.

	Dependent Variable:			
	Satisfaction with Democracy		Support for Decentralization Process	
	ITT	IV	ITT	IV
Encouraged	—0.088*		—0.167***	
	(0.046)		(0.050)	
Attended		—0.397*		—0.803
		(0.214)		(0.272)
Male	—0.022	—0.012	0.064	0.082
	(0.044)	(0.045)	(0.047)	(0.053)
Age	0.003***	0.004**	—0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)
Education	0.020	0.020	0.007	0.003
	(0.016)	(0.016)	(0.017)	(0.019)
Media	0.007	0.0004	0.065*	0.047
	(0.038)	(0.038)	(0.039)	(0.044)
Lagged DV	X	X	X	X
Village FE	X	X	X	X
Observations	1,047	1,047	1,025	1,025
Adjusted R^2	0.413	0.383	0.514	0.380
First Stage F		70.270***		59.320***

p<0.1*; p<0.05**; p<0.01***

Table 4. Effects of VOICE exposure on democratic demand: values and norms.

	Dependent variable:					
	Decentralization Ideal		Tolerance		Right to Criticize	
	ITT	IV	ITT	IV	ITT	IV
Encouraged	0.155**		0.134**		0.104**	
	(0.064)		(0.064)		(0.046)	
Attended		0.772***		0.636**		0.475**
		(0.332)		(0.315)		(0.217)
Male	0.202***	0.185***	0.084	0.076	0.078*	0.059
	(0.061)	(0.063)	(0.061)	(0.063)	(0.043)	(0.045)
Age	0.001	—0.001	—0.005**	—0.007***	0.0003	—0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Education	0.002	0.001	—0.052**	—0.053**	—0.029*	—0.027*
	(0.021)	(0.022)	(0.022)	(0.022)	(0.015)	(0.016)
Media	0.061	0.076	0.036	0.023	0.008	0.0004
	(0.052)	(0.053)	(0.052)	(0.053)	(0.037)	(0.039)
Lagged DV	X	X	X	X	X	X
Village FE	X	X	X	X	X	X
Observations	1,048	1,048	1,070	1,070	1,027	1,027
Adjusted R^2	0.527	0.495	0.516	0.485	0.464	0.420
First Stage F		57.412***		64.640***		67.012***

Note: p<0.1*; p<0.05**; p<0.01***

Selection on the Unobservables: Panel Data

- Panel data offers a wide range of alternative methods for estimating causal effects, 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;
  - Two wave quasi-experimental panel designs: "difference in differences", or "first difference" models
  - Multi-wave, time-varying treatments: "fixed effects"
- These methods are useful for estimating causal effects while controlling for *stable* (time-invariant) unobservables that cause the treatment and control groups to differ at baseline

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	$\Delta Y_{0(D=1-D=0)}$	$\Delta Y_{1\ (D=1-D=0)}$	$(Y_{1(D=1)} - Y_{0(D=1)}) - (Y_{1(D=0)} - Y_{0(D=0)})$

- What we are after is "ρ" – 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*
- This is a very 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")
- Does this solve the "fundamental problem of causal inference"? Sort of !

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

- What is the post-treatment level of Y for the two groups?
  - For the control group: their pre-treatment level of Y, plus a "time" effect that may have changed them ($t_{(D=0)}$). So difference over time =control group time effect.
  - For the treatment group: their pre-treatment level of Y, plus a "time" effect that may have changed them ($t_{(D=1)}$) , plus the "treatment effect" ρ that may have changed them too. So difference over time= treatment group time effect plus treatment effect.
- 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:

(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:

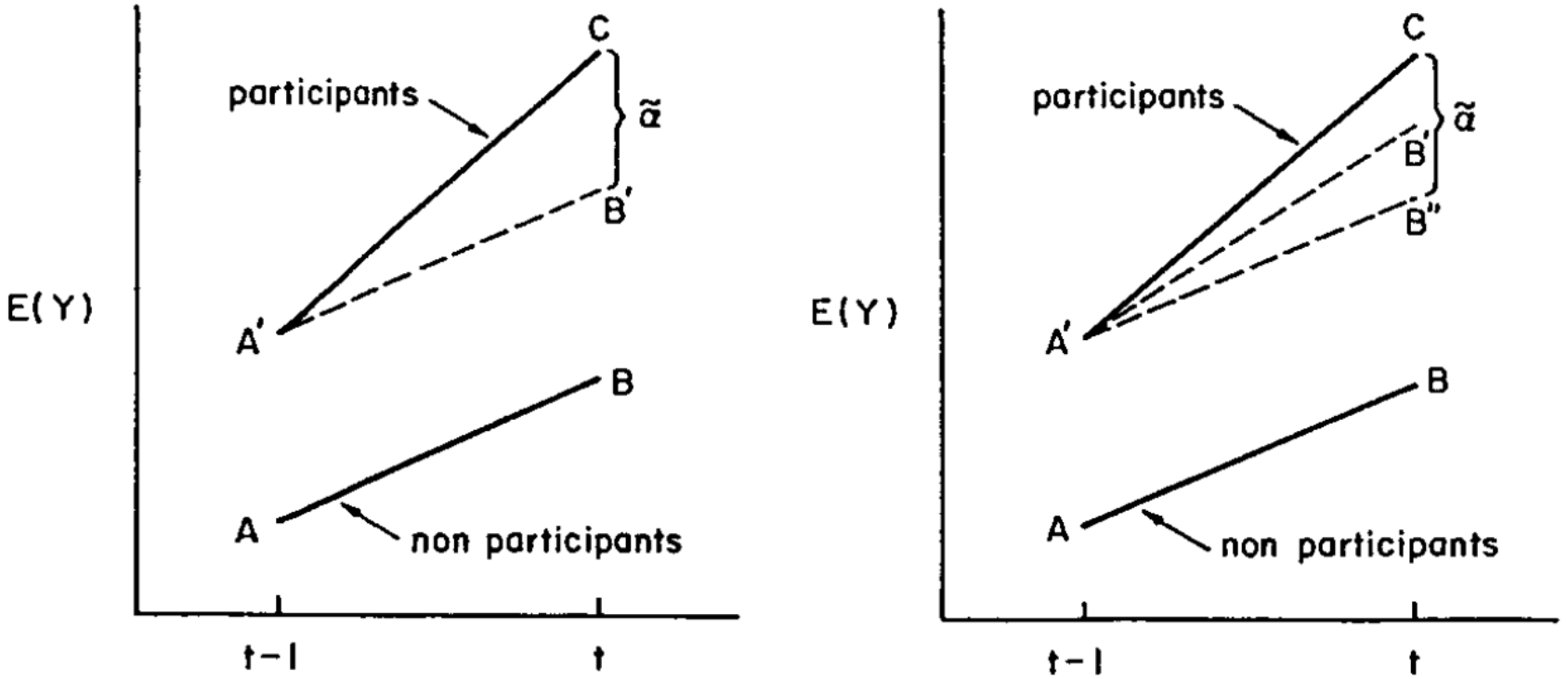
  The observed "difference in differences" will represent the causal effect of the treatment whenever:

$$(4) \quad E(Y^1_{(0)i} - Y^0_{(0)i} \mid D=1) = E(Y^1_{(0)i} - Y^0_{(0)i} \mid D=0)$$

  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**. And since $t_{(D=1)}$ and the causal effect ρ 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 the **"parallel trends" assumption** necessary to identify the causal effect of a difference-in-difference analysis
- 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!
- **This is a great benefit of panel data for causal inference in observational studies!**



- 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 DiD 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 (see PS2701)

- 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*.
  - That is, the treatment group, due to factors that also led them to select into the treatment, may have been changing at a faster rate than the control group, and so would have shown larger changes in Y *in the absence* of treatment. If so, the assumption of our DiD model is invalid and we won't get the causal effect we want.

Regression Estimation of the Two-Wave DiD Model

- Equation (5)

  Time 1: $Y_{i0} = \alpha_0 + \quad + \beta_1 X_{1i0} + \beta_2 X_{2i0} + \cdots \beta_k X_{ki0} + (U_i + \varepsilon_{i0})$

  Time 2: $Y_{i1} = \alpha_1 + \rho D_{i1} + \beta_1 X_{1i1} + \beta_2 X_{2i1} + \cdots \beta_k X_{ki1} + (U_i + \varepsilon_{i1})$

- In the quasi-experimental set up, all D_i at time 1=0
- If U_i is assumed to be stable (time-invariant), by subtraction we arrive at the **first difference, or two-wave DiD model**:

$$(6)\ \Delta Y_i = \Delta\alpha + \rho D_{i1} + \beta_1 \Delta X_{1i} + \ \beta_2 \Delta X_{2i} + \cdots \beta_k \Delta X_{ki} + \Delta\varepsilon_i$$

  or a regression of change in Y against change in the Xs and the indicator for treatment group status

- Simple, but powerful model! The possible confounding influence of **time-invariant** U_i has been eliminated (subject to the parallel trends assumption)

Alternative Model for DiD Estimation

- Another set-up for the two-wave DiD longitudinal model:

$$(7) \quad Y_{it} = a + b_1 D_i + b_2 Time_t + b_3 D_i * Time_t + e_{it}$$

  - 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:

$$(8) \quad \text{Control, Time 0:} \quad Y_{i0} = a + e_{i0}$$

$$\text{Control, Time 1:} \quad Y_{i1} = a + b_2 + e_{i1}$$

$$\text{Treatment, Time 0:} \quad Y_{i0} = a + b_1 + e_{i0}$$

$$\text{Treatment, Time 1:} \quad Y_{i1} = a + b_1 + b_2 + b_3 + e_{i1}$$

- 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 – if we assume that β_2 for treatment and control are equal.
- So β_3 in equations 7-8 is the same as ρ in equations 5-6!
- 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 DiD 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 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

Sønderskov, Dinesen, Finkel and Hansen, "Crime Victimization Increases Turnout: Evidence from Individual-Level Administrative Panel Data" *British Journal of Political Science* (2020)

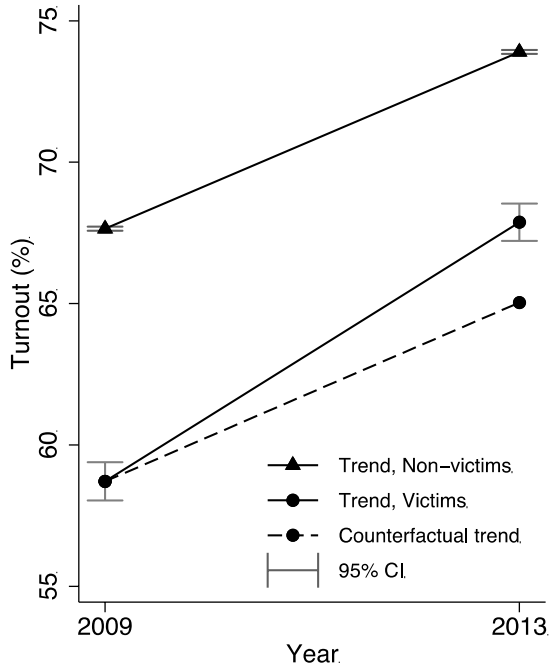


Table 1. The effect of crime victimization on municipal election turnout

Model	1	2	3	4	5
Victimization	Both	Non-violent	Violent	Violent	Violent
Victim	0.005*	−0.002	0.029***	0.029***	0.019*
-within	(2.56)	(−1.14)	(7.44)	(7.44)	(2.10)
Victim				−0.150***	−0.005
-between				(−26.61)	(−0.40)
Constant	0.611***	0.610***	0.606***	−0.058***	−0.303***
	(66.48)	(65.83)	(64.02)	(−18.96)	(−12.59)
Time trend	Yes	Yes	Yes	Yes	Yes
Counter factual trend	Non-victims	Non-victims	Non-victims	Non-victims	Future victims
Time-variant covariates	Yes	Yes	Yes	Yes	Yes
Time-invariant covariates	No	No	No	Yes	Yes
$N_{individuals}$	1,993,359	1,972,752	1,920,847	1,920,847	23,366

Note: *t* statistics in parentheses; two-sided tests. See Appendix A for details on sample sizes for each model, and Appendix D for the full results. * p < 0.05, ** p < 0.01, *** p < 0.001

Extensions

- Multiwave panels for estimating causal effects
- Panel models with lagged dependent variables
- Panel models with differential time-trends and growth curve models
- Instrumental variables in panel analysis to account for effects of time-varying unobservables
- Alternative models for accounting for selection on unobservables
  - Heckman selection models
  - Regression Discontinuity Designs (RDD)

Extensions to Multiwave Panels

$$(1) \quad Y_{it} = a + b_1 X_{i1t} + \ldots b_k X_{ikt} + b_m Z_{im} + e_{it}$$

- Begin with simple pooled model where Y is predicted by two kinds of independent variables: X_k which are time-varying, and Z_m, which are time-invariant. X_{it} has a "t" subscript, not $Z_{i.}$
- Notes:
  - X could be a dichotomous "treatment" variable or a continuous variable – there is no real conceptual difference (though, if we stay within the potential outcomes framework, there are complications in estimating the precise counterfactuals corresponding to each level of a continuous X "treatment")
  - For example, X could be the *level* of democracy of a country (continuous), **or** whether a country transitioned to democracy during that time period (dichotomous). We can estimate these models regardless of this distinction.

$$(1) \quad Y_{it} = a + b_1 X_{i1t} + \ldots b_k X_{ikt} + b_m Z_{im} + e_{it}$$

- Y for a given country-year is function of a common intercept, the regression coefficient (β_1) for time-varying variable 1* X at its value for a given country-year through the regression coefficient (β_k) for the *k*th variable*X at its value for the given country-year, the regression coefficient (β_m) for the *m*th time-invariant variable * Z at its value for the given country, etc., and a country-year error term
- All country-years are pooled together into one regression equation – there is not a separate model for wave 1, for wave 2, etc.

Problems in OLS Estimation of Pooled Model

- Example: Democracy (X) → Repression (Y). An OLS estimation of this relationship will produce a single β for the overall effect of democracy on repression, pooled across country-years
- Problems?
  - **Autocorrelated disturbances** of the error term ε_{it}: if a unit is above the common regression line at time 1, it is also likely to be above the common regression line at time 2, 3, etc. Why? All of the *unmeasured* factors – stable as well as time-varying – that affect the unit over time are lumped into the error term, and these factors will likely be related to one another at times 1, 2, etc.
  - We may also have **heteroskedasticity**: units generally low on X may have little variance around the common regression line, while units generally high on X may have more variance around the common regression line. Or, units at the extremes on X may have little variance on Y compared with units generally in mid-ranges of X. If units have different amounts of error variation generally, and some units are generally higher or lower on X than others, there will be heteroskedasticity in the errors

- So we know that OLS might be problematic, because the OLS assumption is that:

$$E(\sigma_i^2) = \sigma^2 \quad \text{for all levels of X} \quad (\text{"homoskedasticity"})$$

$$E(\sigma_{ij}) = 0 \quad \text{for all observations i and j} \quad (\text{"non-autocorrelation"})$$

- Both of these assumptions are likely to be violated because OLS wants to treat all observations as independent, and since the observations here are on the *same units* at different points in time, they are not truly independent. This is another way of saying that because the observations are ***clustered*** by unit, important OLS assumptions are likely not to hold.
- From the violations of the error term assumptions discussed so far, we can say that OLS at minimum is likely to produce *inefficient* estimates of the β compared to other estimators

Unobserved Heterogeneity

- But problem goes deeper, because of the "U" term we have discussed previously, i.e., *unobserved heterogeneity* that may be correlated with observed independent variables
- That is, one of the reasons for the autocorrelation itself is that stable, unobserved factor or factors that are unique to a given country (unit) make that country (unit) generally higher or lower than the average country (unit).
  - For repression/democracy example, it may be cultural or historical factors, ethnic separatism, religious traditions, size of the military, alliances with dictatorships and democracies, all of which may play a role in pushing countries generally higher or lower on repression.
  - If these variables can be measured, then of course we want to bring them in to the analysis directly.
  - **AS WE KNOW, HOWEVER, WE ARE NEVER (OR NEARLY NEVER) ABLE TO MEASURE AND INCLUDE ALL RELEVANT FACTORS THAT INFLUENCE THE DEPENDENT VARIABLE**. If we cannot, they become part of the error term.

$$(2) \quad Y_{it} = a + b_1 X_{i1t} + \ldots b_k X_{ikt} + b_m Z_{im} + U_i + e_{it}$$

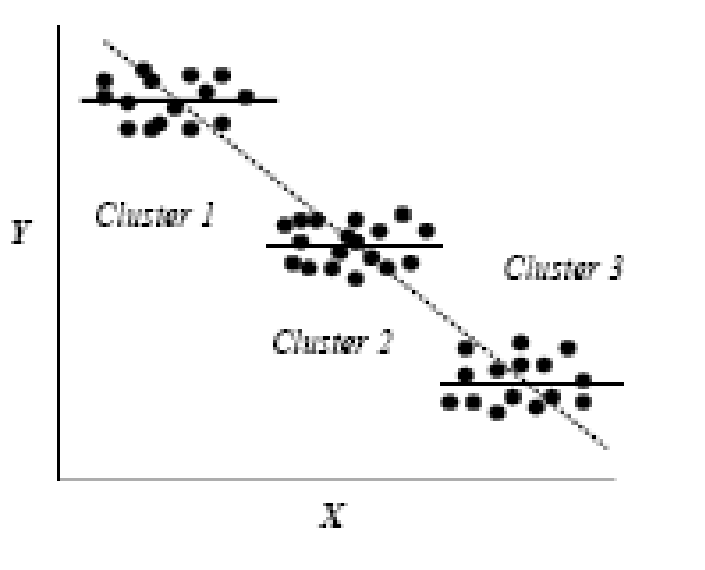
- With U_i representing all of the "***unobserved heterogeneity***," or the unobserved *stable* factors in case or unit *i*. (Unobserved "unstable" factors are still in the error term ε).
- The error term in this model is now composed of two parts: a unit-level effect that does not vary across time (U_i) and an idiosyncratic error term that varies across units and across time (ε_{ij}). This composite error term ($U_i + \varepsilon_{ij}$) decidedly does not conform to OLS assumptions.
- The U term is called a "unit effect," a "permanent effect," or a "fixed effect", though the last term is confusing because that is also one of the ways of dealing with it (the so-called "fixed effect model"). Much of panel data econometrics is designed to estimate the β efficiently and without bias in the face of the unit effects that induce problems in the error term.

Implications of U

- The intercept (α) is no longer common to all units. In fact, every case has its own intercept , ($\alpha + U_i$).
  - E.g. Turkmenistan is generally high on repression due to a large positive U_i; Costa Rica might be generally low due to low U_i, or stable unmeasured variables that make it lower at all points in time.
  - So unobserved heterogeneity leads to the violation of the common intercept assumption of the Pooled Model also, as well as inducing autocorrelation in the disturbances
  - OLS estimation is inefficient, because it does not take into account that some of variance in Y is due to the common unit effects from each group. Once we control for that (through estimation of the individual intercepts), we would have lower variance around the individual regression lines. We may also have little or no autocorrelation left, if all of the temporal dependence is due to U_i.

- Problems more severe if U_i term is related to X variables that are included. This leads to biased estimates of the β in (2).
  - Why? Because the included Xs will be related to the composite error term ($U_i + \varepsilon_{ij}$)! Therefore will again violate the OLS assumption that E(Xε)=0, just as we saw with models that include reciprocal causality and/or measurement error. In this case the "**endogeneity"** problem is induced because of *omitted variable* bias in the form of X being related to U_i.
  - This is another type of the "selection on the unobservables" problem in the counterfactual causal inference framework
  - Example: Countries with generally larger militaries (as percent of GDP) are perhaps less likely to be democratic, and perhaps more likely to repress their citizenry. If so, democracy *per se* may not be related to repression at all. It is only that large militaries are related (negatively) to democracy and (positively) to repression, but since you haven't observed this variable, there looks to be a spurious relationship between repression and democracy is. So the unmeasured variable U_i is responsible for the repression-democracy relationship, and failure to take this into account leads to BIAS in estimation of the effect of democracy on repression. Controlling for U_i (if we could) would show us that the true democracy β would be 0.

Example of Cluster Bias, or Bias Caused by Unobserved Heterogeneity



X: Democracy
Y: Repression
U: Size of Military (among other things)
X→-Y overall in pooled model, **BUT**
U->+ Repression (Y), U->-Democracy (X), but controlling for U (within clusters), X has no effect on Y

- The pooled model shows a negative effect of X→Y
- But cluster (unit) 1 is generally high on Y, cluster 2 generally middle, cluster 3 generally low on Y, and:
- Whatever is causing the clusters to differ on Y appears to be related to X as well (cluster 1 is low on X, cluster 2 middle, and cluster 3 high on X)
- "Within" each cluster, there is NO X→Y relationship at all!
- So failing to consider the "unit" effect on Y and its possible correlation with X results in erroneous inferences!

$$(2) \quad Y_{it} = a + b_1 X_{i1t} + \ldots b_k X_{ikt} + b_2 Z_i + U_i + e_{it}$$

- Notes:
  - Model (2) with endogenous X is very difficult to estimate with cross-sectional data! The literal model of (2) is impossible to estimate, since the unobserved variable U_i is folded into ε_{it} and there is no way with cross-sectional data to produce an "estimate" of U_i or to unpack its independent effects. As noted, you can estimate the β in (2) with good instrumental variables, though, as we have also noted, these are very difficult to find.
  - It is also possible that not only the *intercept* differs across units, but also the *slope* for democracy (or other variables). What if democracy strongly affects repression in country 1, less so in country 2, etc? This leads to "random coefficient" models that we "won't have time to talk about" (!)

Fixed Effects Model

- The basic idea of "fixed effects": if the intercepts differ for each country in equation (2), then let's include a dummy variable for each case (minus one baseline case), and we end up with N-1 intercepts, which, when added to the overall α, give us N different "starting points" or "average" values of Y for each unit. We then estimate the effects of the other Xs, controlling for the unit-level starting point or average value. This approach is called the "LSDV" estimator, for "Least Squares Dummy Variables."

$$(3) \quad Y_{it} = a + b_1 X_{1it} + \ldots b_k X_{ikt} + b_m Z_i + c_1 D_1 + c_2 D_2 + c_3 D_3 + \ldots c_{n-1} D_{n-1} + e_{it}$$

where D_1 is a dummy variable for unit 1, D_2 is a dummy variable for unit 2, until D_{n-1} is a dummy variable for unit n-1.

($D_{1,2,3}$ etc., is **not** to be confused with the "treatment variable" D from earlier)!

- So the intercept for unit 1 is (α+c1), the intercept for unit 2 is (α+c2), for unit 3 it is (α+c3), and so on until the nth-1 unit which has an intercept of (α+cn-1). The nth unit's dummy variable is not included, so it will have an intercept of α.
- Thus the dummy variable's regression coefficient is the estimate of U_i! (Technically, it is the estimate of U_i *plus the effect of all stable observables*, which cannot be distinguished from the dummy effect)
- NOTE: This shows why we cannot estimate this model with cross-sectional data. You cannot add a dummy variable for each case, as there are not enough degrees of freedom nor unique pieces of information available to estimate such an effect, independent of the other variables in the model!! (TRY THIS AT HOME – THERE IS NO WAY TO SEPARATE THE ERROR TERM AND THE UNIT-LEVEL INTERCEPT!)
- We call the LSDV (and the Fixed Effects) estimator **"WITHIN ESTIMATORS"** of the βs, because they completely control for variation **between** units through the dummy variables, and only use "within-unit variation" in X to estimate the βs. This will be seen more clearly below.

- Big Problem with LSDV, though: Computationally, there is the addition of possibly thousands of dummy variables in the model, which may be more than even STATA with dual core processors can handle!
- In some maximum likelihood panel models, e.g., those for dichotomous variables that we will consider later in the course, it is not possible to just add more and more dummy variables without affecting the consistency of the estimates (this is called the "incidental parameters" problem in ML estimation, see famous paper by Neyman and Scott (1948) which shows inconsistencies as the number of nuisance parameters (like dummy variables for each case) increases relative to number of observations).
- To get around this problem, we do the following trick on the next slide to arrive at what is called the "Fixed Effects" Model
- The coefficients for the effects of the Xs will be *identical* to what you would have obtained with LSDV methods

- Start with another version of (2), this time expressed in terms of the *means* of all variables:

$$(4) \quad \bar{Y}_i = a + b_1\bar{X}_{1i} + b_2\bar{X}_{2i} + \ldots b_k\bar{X}_{ki} + b_m\bar{Z}_{mi} + \bar{U}_i + \bar{e}_i$$

- NOTE: THIS EQUATION IS ALWAYS TRUE IN LINEAR REGRESSION
- If use OLS on this, we obtain the so-called "BETWEEN ESTIMATOR" of the βs because it completely controls for variation within the units through the averaging process, and only uses ***between-unit*** variation in the Xs to estimate the Bs.
- Now subtract (4) from (2) and you get:

$$Y_{it} - \bar{Y}_i = (a - a) + b_1(X_{1it} - \bar{X}_{1i}) + \ldots b_k(X_{ikt} - \bar{X}_{ik}) + b_m(Z_{im} - \bar{Z}_{im}) + (U_i - \bar{U}_i) + (e_{it} - \bar{e}_i)$$

- Or what is known as the FIXED EFFECTS (FE) model:

$$(5) \quad Y_{it} - \bar{Y}_i = b_1(X_{1it} - \bar{X}_{1i}) + \ldots b_k(X_{ikt} - \bar{X}_{ik}) + e^*_{it}$$

$$(5) \quad Y_{it} - \bar{Y}_{i} = b_1(X_{1it} - \bar{X}_{1i}) + \ldots b_k(X_{ikt} - \bar{X}_{ik}) + e^{*}_{it}$$

- The FE Model:
  - An OLS Regression of "Demeaned-Y" against "Demeaned X"
  - Eliminates the U_i term from consideration through the "demeaning" process. They have been "swept out" of the equation!
  - All that is left is "pure" error ε_{it}^{*}, plus the variation of X and Y around their unit-level means
  - Another view of the "within estimator" that LSDV or FE represents: we are dealing only with variation *within* each case over time – as X changes from its unit-specific mean, does Y change from its unit-specific mean?
  - The average *level* of X and Y have been subtracted out of the model!
  - So with FE we can estimate the effects of X on Y with longitudinal data, controlling for the potentially biasing effects of unmeasured stable variables! (So long as the assumptions of the method hold).

Notes/Issues with FE Models

1. One may test for the statistical significance of the unit effects, jointly considered. That is, do all of the unit effects, taken together, explain a significantly greater amount of variation in Y than a model with a common intercept? So a test of overall unit effect significance is a test of the difference in R^2 between a "constrained" equation, where all $U_i = 0$, and an "unconstrained" LSDV equation. This is an F* test:

$$F^* = \frac{\frac{R^2(unconstrained) - R^2(constrained)}{N-1}}{\frac{1 - R^2(unconstrained)}{NT - N - k}}$$

where NT is the total number of observations, N is the number of units, and k is the number of regressors in the model.

2. When estimating the FE model (5) with OLS, the standard errors need to be adjusted to reflect the fact that you lose N degrees of freedom in calculating the unit means. So the FE model has NT (total observations) – N (units) – k (regressors) degrees of freedom for the estimates of the standard errors (as the denominator in the F* denominator above shows). STATA/R automatically make this adjustment in their calculations. This also means that standard errors in FE are typically larger than in OLS, since we have lost the *df* which figures in s.e. calculations
3. It is easy to recover an intercept in the FE model (remember from (5) that the α drops out from the differencing process). You just add the overall sample mean (the "grand mean") for Y and X to the "demeaned" Y and "demeaned" X and run the FE regression on those new variables. This is what STATA does, which is why you get an estimate for α in the FE model.

4. The main drawback of FE models: Look at what happens to all other stable or time-invariant independent variables **Z** in the De-Meaning process. They also drop out!!! They are exactly like Unit Effects—stable variables at the unit level. (Same thing happened in DiD model as we saw last session).

Moral: WITHIN estimators cannot provide separate estimates of Stable Observed and Stable Unobserved Variables. (Variables that change very little over time are also problematic -- little change in X means unreliable estimates).

This can be a *huge* problem when such variables are of prime theoretical interest (as in much comparative work interested in institutions and/or political or economic structural factors. But TANSTAAFL!!!! *(at least until the section on "compromise" or "hybrid" models!). We give up the ability to say much about stable variables, in return we get estimates that control for potential endogeneity due to stable unobservables.

*** "There Ain't No Such Thing As A Free Lunch"!!!**

5. R-squared is a tricky business with pooled panel models in general, because there are many kinds of variance you can think about explaining. There is the "within" R-squared (controlling for "between" effects), the "between" R-squared (controlling for "within" effects), and the "total" R-squared, some kind of combination of the two.

We can use the idea of R^2 being the squared correlation of some predicted value of Y --($\hat{Y}$)-- with the actual Y to explore these ideas.

- Total R^2 : Get the predicted Y from $\Sigma(XB)$ from the FE estimation, and correlate this with actual Y.
- Between R^2: Generate the AVERAGE of $\Sigma(XB)$ for each unit across all time periods, and correlate this with AVERAGE Y across all time periods. This averages out all of the "within" variation in $\hat{Y}$ and in Y.
- Within R^2: The R^2 you would obtain from direct OLS estimation of equation (5). Practically, it is the squared correlation of demeaned predicted $\hat{Y}$ (from demeaned $\Sigma(XB)$) with actual demeaned Y. ***This is the preferred R-squared to report***.

6. We can also take possible heteroskedasticity of the error terms into account by estimating so-called "ROBUST" standard errors which, in the panel context, also take further into account the potential error term problems produced by unit-level clustering (the so-called "cluster-robust" standard errors). This is done with the **VCE(cluster clustername)** option in STATA.

$$\text{OLS variance of } \beta = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{where } \sigma^2 = \frac{\sum_{i}^{n}(\varepsilon_i^2)}{N-k}$$

- With heteroskedasticity, this variance no longer constant for all X. White's "Heteroskedastic-consistent" standard errors for the non-panel case:

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2\sigma_i^2}{(\sum_{i=1}^{n}(X_i - \bar{X})^2)^2}$$

- Each level of X_i has its own σ^2. In practice, we do not observe the σ^2 so we use the OLS residuals at each point on X as the best guess. The square root of this quantity is called the "ROBUST" standard error

- Longitudinal extension: average these across all the units or "clusters" to arrive at "Clustered" Heteroskedastic-Consistent standard errors:

$$\frac{\sum_{c=1}^{C}\sum_{i=1}^{n}(X_i - \bar{X})^2 s_i^2}{(\sum_{i=1}^{n}(X_i - \bar{X})^2)^2}$$

  where c is an individual unit (cluster) and C is the total number of units (clusters). Square root of this is the standard error used.

- It can be seen that the difference between the clustered and unclustered version of the ROBUST estimate is that, in the clustered version, the numerator represents the sum of C **averages** of the products of the Xs and the errors, while in the unclustered version, the numerator is simply the total product of all the Xs and the errors. So the clustered version takes into account the non-independent nature of the observations and (so to speak) aggregates the estimate of heteroskedasticity by cluster.

- The actual values of the unit effects are usually not that substantively interesting, but in some instances you may want to examine them (especially if your units are countries, states, etc. as opposed to individuals). If so, you can calculate them easily as:

$$\hat{u}_i = \bar{Y}_i - \sum_{k=1}^{K} b_k \bar{X}_{ik}$$

(remember to include the overall constant β_0 in this calculation)

- And you can also obtain an estimate of the proportion of the overall composite error variance that is made up of "unit" variance versus "pure" idiosyncratic variance. This is the *Rho* statistic (sometimes also called the "intra-class correlation coefficient or ICC") that tells you how much overall error is unit-time-specific and how much is simply unit-specific:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}$$

Time and the "Two-Way Fixed Effects" Model

- It is very easy to extend the FE model to include TIME effects as well. You simply add dummy variables representing each time point (minus one baseline category) to the original longitudinal model

$$(6)\quad Y_{it} = a + b_1 X_{1it} + \ldots b_k X_{ikt} + b_m Z_{mi} + t_1 T_1 + t_2 T_2 + \ldots t_{t-1} T_{t-1} + U_i + e_{it}$$

- Subtracting the "Between" equation from this results in a "demeaned" and "detimed" model where the effects of time are controlled so that you do not confuse effects of X with the general effects of time or particular time periods that might influence *all* cases (i.e., these dummies capture the general effects of time, even for cases that never change on X)
- Following this logic, the "t" effects must be assumed to be the same for individuals at all levels of X (the "treatment") would have been in the absence of treatment, just like in the DiD model!

- So the independent variable in this model is the deviation of X for case *i* at time *t* from its unit-specific mean ***and*** from the overall sample's mean at time *t*. If we add the "grand mean" so as to obtain an FE intercept, that means the "demeaned"/"detimed" variables would be calculated as follows:

$$Y_{it}^{*} = Y_{it} - \bar{Y}_{i} - \bar{Y}_{t} + \bar{Y}$$

$$X_{it}^{*} = X_{it} - \bar{X}_{i} - \bar{X}_{t} + \bar{X}$$

- Estimate by regressing Y_{it}* against X_{it}* -- or by doing traditional FE and adding the time dummies. The latter is easier in STATA.
- **This model is called the "Two-Way Fixed Effects" Model and is almost *always* preferred over the simple one**
- NOTE: These time dummies are atheoretical!!! You are just controlling for any possible period effects or shocks that may affect all units at a given point in time.
- So think carefully about time effects – if you think there is a trend in the process that affects everyone, e.g., include TIME as a counter or trend variable.

Extensions: Time-Varying Effects of Covariates

- In the FE model, time-invariant variables drop out, but *only* so long as their effects are assumed to be constant over time. In FE, this assumption can be relaxed by including an *interaction* between Z and the time dummies (or a time trend):

$$(7)\ Y_{it} = a + b_1 X_{1it} + \ldots b_k X_{ikt} + b_m Z_{mi} + t_1 T_1 + b_{mt1} Z_m * T_1 + t_2 T_2 + \ldots t_{t-1} T_{t-1} + U_i + e_{it}$$

- Even though Z by itself drops out of the estimation, Z*T is not time-invariant, so will not drop out, so the coefficient β_{mt} will pick up the differential effect of Z at time period T on demeaned-Y compared to its "average" effect
- Same logic holds for including an interaction between time-varying X with a time trend or (more awkwardly in long panels) with each time dummy to pick up differential effects of demeaned X on demeaned Y at a particular point in time, or in a linear increasing fashion over time

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		Total Respondents	
			Follow-up from Wave 1 Only (C)	Follow-up from Waves 1 and 2 (D)	All Two- or Three-Wave (B+C)	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*

Example of 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)

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.

Strengths and Weaknesses of the FE Approach

- Strengths
  - Eliminates consideration of unobserved U_i through differencing or partialling, and thus:
  - Estimates effects of time-varying X while controlling for possible correlation with U_i. Solves the endogeneity problem caused by *stable unobservables* that are correlated with the included Xs (provided the assumption of equal "no treatment" time trends for units at all levels of X (or D) is satisfied)
  - Close relationship of FE with treatment effects and quasi-experimental models for estimating causal effects

- Weaknesses
  - Impossible to say anything about effects of *time-invariant* variables
  - Focusing solely on *within* variation ignores the question of *why* some units are generally lower or higher than others, i.e., *between* variation. FE doesn't model between variation, it just takes it as given and "sweeps" it out of consideration altogether. But modeling between variation might be of theoretical interest in its own right.
  - FE with few time points can estimate U_i unreliably; a few random high or low values on Y for a given unit will look like "U_i" and not random noise. FE uses whatever is in our sample data for each unit, perhaps not the most efficient way to estimate a "unit effect"
  - We lose *N* degrees of freedom in calculating FE models – in small T studies, this affects efficiency of estimates and produces larger standard errors than (perhaps) necessary
  - FE cannot correct for biases due *time-varying* unobservables that affect the X and Y

The Random Effects Model

- Alternative to FE/FD: "Random Effects" (RE). Gets around some of these problems, but has its own set of possibly problematic assumptions as well. (**"TANSTAAFL!")**
- Go back to original model of heterogeneity:

$$(1) \quad Y_{it} = a + b_1 X_{1it} + b_2 X_{2it} + b_3 X_{3it} + \ldots b_k X_{ikt} + U_i + e_{it}$$

- Look at composite error term: $(U_i + \varepsilon_{it})$
  - RE says: let's estimate the model by treating **both** components of the error term as arising from (independent) random processes. This is the way ε_{it} is usually modeled. What is new here is treating U_i the same way. Instead of being a "fixed" quantity at the level that is produced by our sample observations, we treat U_i as a second normally distributed variable so that each unit's U_i is the result of a random draw from this second distribution.
  - Some units higher on Y generally because of a randomly drawn high U_i , some units lower on Y generally because of a randomly drawn low U_i
  - This is why the RE model is sometimes called the "**Random Intercept**" Model – the draw from the second distribution, added to the grand mean of Y, gives each unit its observed intercept.

Assumptions of RE

1. The two components of the composite error, U_i and ε_{it} are independent, i.e. $E(U\varepsilon)=0$
2. The variances of both U_i (σ^2_u) and ε_{it} (σ^2_ε) are constant for all X (no heteroskedasticity)
3. The idiosyncratic residuals ε_{it} at one point in time are not related to their value at another point in time (no autocorrelation in ε_{it}).

So far, so good, these three are relatively unproblematic.

4. Both U_i and ε_{it} are unrelated to the X_{ik}, i.e. $E(XU)=E(X\varepsilon)=0$

Yes, you read that right!! In order to identify and estimate the β in the RE model with *two* separate error terms, we need to treat them as *unrelated to the observed independent variables*. This is the usual OLS assumption for ε, and now we have to extend it to U as well. Otherwise we can't estimate the separate effects of X and the composite error.

- If assumption (4) is true, then RE is **definitely** the best estimator available. It is the most efficient, since (as we will see) we are only adding one additional parameter to the estimation --- the variance of U_i (σ_u) --- instead of the *N-1* new parameters in FE/FD. But if this assumption is **not** true, then random effects gives wrong answers, as the RE estimator will be biased (technically, "inconsistent": biased even when $N \rightarrow \infty$)
- In fact the potential violation of this assumption is the reason that many panel analyses are done in the first place!!!!
- Thus, many people dismiss RE out of hand, and believe FE is the way to go. However, let us see what the RE estimator entails, how it works, what its advantages may be, and then how we might make use of it, if not in its "pure" form, then in some modified form to add to our toolkit of longitudinal methods.

Estimation of RE Models

- Problem in estimating (1)? The composite error term ($U_i + \varepsilon_{it}$) has a complicated structure that is no longer independent over time within cases because of the presence of U_i
- The variance of the composite error, given assumptions (1)-(3) above, is: $\sigma^2_u + \sigma^2_e$. The covariance of the composite error is: σ^2_u for all time periods.
- So the variance-covariance matrix of the error term, given RE assumptions, is (for a 4 wave panel as an example):

(2)
$$\begin{bmatrix} \sigma_u^2 + \sigma_e^2 & & & \\ \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 + \sigma_e^2 \end{bmatrix}$$

- How do we estimate a model such as equation (1) that has an error structure like matrix (2)?
- Several ways, but the simplest is to estimate via **Generalized Least Squares (GLS),** which involves ***weighting*** the equation by a factor that will transform the problematic error term (2) into a variant of the unproblematic error term (3), so that OLS can be used on the *weighted* or *transformed* model.
  - Remember this – WLS/GLS from Heteroskedasticity session?
  - Another example of GLS is in time-series analysis, where one might weight the data by ϱ, the autocorrelation parameter for the ε_{it}, and then use OLS on the weighted data
- GLS proceeds by weighting the data by the *inverse* of the error variance-covariance matrix to ensure that the weighted equation has a normal structure with common variance on the diagonals and zero covariances on the off-diagonals. Then OLS is used on the weighted equation.

GLS Random Effects Estimation

- Needed: estimates of the two variance terms $\sigma^2_u + \sigma^2_e$. If we could obtain those estimates, we can weight or transform equation (1) in the following way and then use OLS to estimate the effects:

(4) $$Y_{it} - q\bar{Y}_i = (a - qa) + b_1(X_{1it} - q\bar{X}_{1i}) + b_2(X_{2it} - q\bar{X}_{2i}) + b_k(X_{ikt} - q\bar{X}_{ik}) + (U_{it} - q\bar{U}_i + e_{it} - q\bar{e}_i)$$

with:

(5) $$q = 1 - \left(\frac{s_e}{\sqrt{T s_u^2 + s_e^2}} \right)$$

- If the observed Y and X in the model are transformed/weighted by equation 5's θ ("Theta") then the resulting error term in (4) will be OLS-ready, i.e. with constant variance on the diagonals and zero covariance on the off-diagonals.
- We don't know the population values of these two error variances. We need to *estimate* them from our data, which is why this application is called **"Feasible Generalized Least Squares" (FGLS)** and not the "real" thing. It is a bit more imprecise, but this is taken care of in various adjustments of degrees of freedom and thus the standard errors of the resulting coefficients.
- This procedure yields the Random Effects (RE) parameter estimates

RE versus FE: The Hausman Test

- So, should you use FE or RE? Clearly, the choice is based on whether the RE assumptions (1) through (4) above are satisfied. If they are, the RE is more efficient, as stated earlier. If they are not, then RE will be biased and *inefficient*, and this is usually taken to mean that FE will be preferred.
- A test developed by Hausman exists to assist in this choice. The intuition of the Hausman test: If the assumptions of RE hold, then FE and RE are two different ways to get the correct estimates, but RE is more efficient. If the assumptions do not hold, then RE will be "inconsistent" (biased even as sample size gets larger and larger), and only FE will be "consistent". So we should see similar estimates between RE and FE if the RE assumptions hold, and different ones if they don't.
- So the Hausman test is:

$$(6) \qquad \frac{\beta_{FE} - \beta_{GLS-RE}}{\text{var}(\beta_{FE}) - \text{var}(\beta_{GLS-RE})}$$

with the test statistic being distributed as χ^2 with degrees of freedom equal to the number of time-varying independent variables, under the null hypothesis that:

$\beta_{FE} = \beta_{GLS\text{-}RE}$.

- Rejecting H_0 is usually taken to mean that the RE assumptions – in particular, no correlation between X and U – do not hold, and that therefore FE is preferred. If not, we would have seen similar estimates from RE and FE, given sampling error. But this is not exactly true….

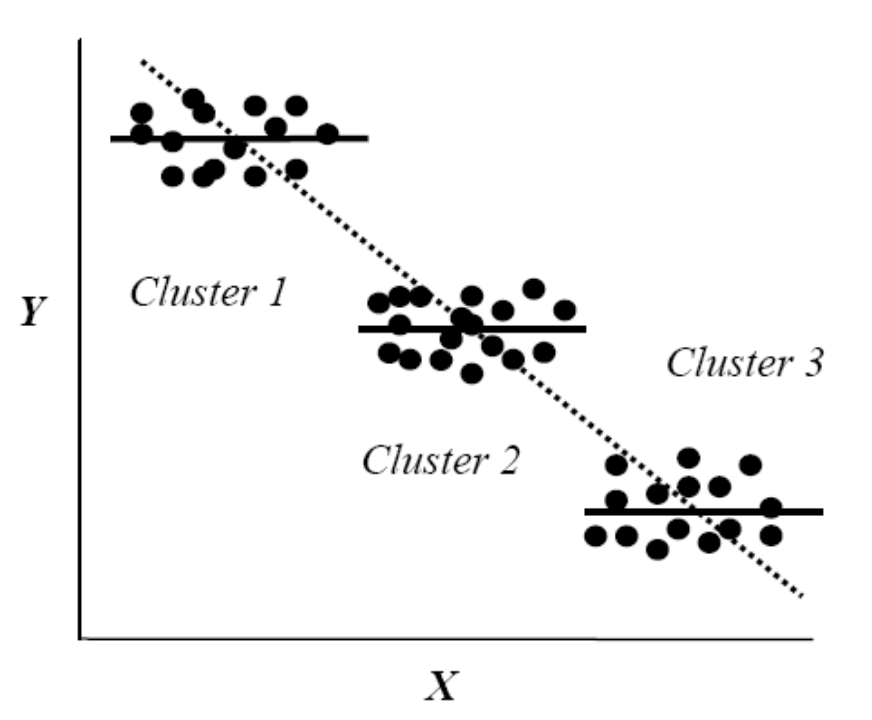
The Random Effects-Hybrid Model

- There is a compromise model: The "RE-Hybrid Model discussed by Bell&Jones (2015) and others, beginning in ecomomics with Mundlak (1978). It is extremely simple and might be labeled under the category: "what is all the fuss about?"
- The idea here starts with the notion that the possible covariation of time-varying Xs and the U_i is what messes up RE. But this possible covariation is usually just the result of model misspecification --- something in the U_i term is related to something in the X that we need to account for, and RE (at present) cannot account for it because of its assumption that E(XU)=0. But we can bring the covariation between X and U_i into the model indirectly, by including the **mean of X** as an additional independent variable in (4):

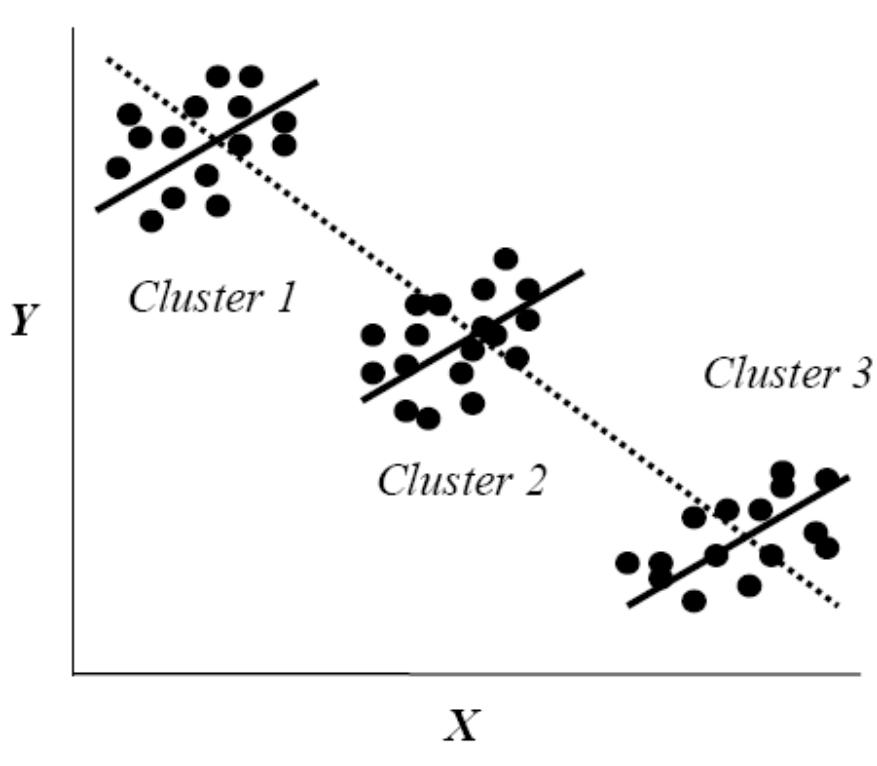
$$(7) \quad Y_{it} = a + b_1 X_{1it} + b_2 \bar{X}_{1i} + U_i + e_{it}$$

- Whatever covariation between X and U that may exist is now accounted for; if units that are generally high (low) on X also have high (low) U terms, then the mean of X in the model will pick this up. The effect of "regular" X can now be estimated, controlling for this possible confounding problem. Sounds like FE!!

Examples: X-bar picks up the X-U correlation: Clusters low on X(bar) are high on U, clusters high on X(bar) are low on U. Controlling for X-bar allows estimation of the "within" effect of X, just like FE did!



Null "within" cluster effect, negative "between" cluster effect



Positive "within" cluster effect, negative "between" cluster effect

- So we now have a U_i, X_{it} and X-bar in (7), and the β for the time-varying values of X are thus estimated in RE while controlling for the possibly confounding correlation between X and U_i. So we can use RE for all the reasons we like RE:
  - only one lost df and thus more efficient estimates
  - the ability to model the intercepts and try to account for why some units are higher or lower than others
  - the ability to include other TIV in the model as predictors
- As we will see, the RE model is also the baseline model for the multilevel/hierarchical school or framework for longitudinal analysis. This simple little correction shows that we can use this framework and still not sacrifice some of the advantages of FE
- **HEY, MAYBE THERE IS A FREE LUNCH AFTER ALL?**
  - Not completely. Need to assume that (residualized) U is unrelated to all Z, so effects of Z may be overestimated (as Z takes any correlated effects it may have with U for itself). This is not a **huge** drawback but still a drawback.

- Can examine the "hybrid" model more closely:

$$(8) \quad Y_{it} = a + b_1 X_{1it} + b_2 \bar{X}_{1i} + U_i + e_{it}$$

- Can write this model in terms of the mean-deviations in X too:

$$(9) \quad Y_{it} = a + b_1 (X_{1it} - \bar{X}_{1i}) + (b_1 + b_2) \bar{X}_{1i} + U_i + e_{it}$$

$$(10) \quad Y_{it} = a + b_1 (X_{1it} - \bar{X}_{1i}) + b_3^* \bar{X}_{1i} + U_i + e_{it}$$

- With $b_3^* = \beta_1 + \beta_2$ from equation (8). Equation (10) expresses the model in terms of the mean of X, and the deviation of X_{it} from the mean of X.
- We can see from this expression that the Hybrid coefficient β_1 is going to give you the same value as the FE estimate, and coefficient β_3 is going to give you the BE estimate (the "between" estimate) of the causal effect of X on Y. In other words, the coefficient for β_1 will give you the effect of a **within-unit** change on Y_{it} (i.e. changing a given case by one unit over time), and the coefficient for β_3 is going to give you the effect of a **between-unit** change on Y_{it} (i.e. changing a given case into another case that is, on average, one unit higher).

- Will these two estimates generally be the same?
- Traditional RE models (see equation 1) assume that the answer is yes. That is, they assume that $\beta_1 = \beta^*_3$ so that the effect of X-bar on Y is zero. Thus, we can say that the traditional RE formulation says that the FE and BE estimates of the effect of X (or the "within" and "between" effects) are equal!!!
- In fact, we can conduct a statistical test of the equality of these coefficients in STATA using the "test" command after running an RE version of equation (10):

  test $\beta_1 = \beta^*_3$

- **THIS IS EXACTLY THE SAME AS THE HAUSMAN TEST FOR FE VERSUS RE DESCRIBED ABOVE!!!!!**

- So the traditional RE assumption is the same as saying that the *within-person (unit)* and *between-person(unit)* effects are the same. If this assumption is wrong, then traditional RE is wrong.

- When will this be wrong? When you can imagine that comparing units that differ from one another by one unit on X will lead to different effects on Y from the effect of changing a single unit at time 1 by one unit to time 2. (See earlier graphs on cluster bias).
- Examples: Typing Speed and Error Rates; Exercise and Mortality
- Social Science Example:
  - Persons XYZ are crime victims, Persons ABC are not. Persons XYZ may have *lower* likelihoods of voting than persons ABC, i.e. the **between** effect of "VICTIM" will be negative. Victims are the kinds of people who tend to vote less (they are more socially isolated, less educated, live in more crime-ridden areas, etc.)
  - But Person X is now victimized. She (and persons B and C after their respective vicitimizations) may be subsequently motivated to vote due to "post-traumatic growth" or the desire to change crime policies. The "**within**" effect of VICTIM might therefore be positive. [Sonderskov et al, BJPS 2020 from the DiD presentation!]
- Or:
  - Gelman's (cross-sectional) work on US voting behavior: Persons living in rich states are more likely to vote Democratic than persons living in poor states (i.e. the "**between** effect" of INCOME on the Republican vote is *negative*). But rich people in a given state are more likely to vote Republican (i.e "the **within** effect" of INCOME on the Republican vote is *positive*).

Moral of the Story

- Rejecting the null hypothesis, or rejecting RE assumptions in the Hausman test either means that FE is valid, *or* that there is some omitted effect of Xbar on Y, independent of the effect of (X_i-Xbar). As Skrondal and Rabe-Hesketh state in their book *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models* (Chapman and Hall 2004, p. 53):
  - *"Some economists believe that a significant Hausman test implies that the random intercept model must be abandoned in favor of a fixed effects model. However, this is misguided since β_1 can be estimated without bias as long as the cluster mean is included as a covariate in addition to X_{it}"*
- The model with both Xbar and (X_i-Xbar) as independent variables allows RE estimation in the context of potential XU_i correlation, *and* it allows interpretation of possible differences in the *within* and *between* estimates, which may have substantive implications in a given analysis.
- Staying within the RE framework has other advantages, e.g. allowing estimation of stable observables, and allowing estimation of random coefficient models mentioned earlier. So the "hybrid model" could very well be the best model for (relatively) short-T panel studies.