Fixed Effects and First Difference Models for Multi-Wave Panel Data

PS 2701-2019

Week 2

Professor Steven Finkel

Extension to Multi-Wave Panels

- With more than two waves of observation, options increase for estimating models that control for stable unobservables:
- Three basic choices:
  - "Fixed Effects" Model (FE)
  - "First Differences" Model (FD), extending our treatment of two wave QE models from last session
  - "Random Effects" Model (RE)
- More recently, compromise models attempting to integrate FE and RE ideas
  - Plümper-Troeger Fixed Effects Variance Decomposition (FEVD)
  - Hausman-Taylor Model
  - Bell-Jones Hybrid Model (RE-Hybrid)
- Nearly all of these models are available with two wave data as well, though (in political science) difference models tend to dominate, given their strong linkage to QE/experimental designs

"Long" versus "Wide" Data Structures

- Multi-wave panel models, including FE, FD and others we will consider, are usually estimated with what is referred to as **"long"** format data
- Data from earlier two-wave example was in **"wide"** format, with rows representing units and columns representing variables. For multi-wave longitudinal data, wide format data would look like this:

	Y(Wave 1)	X(Wave 1)	Y(Wave 2)	X(Wave 2)	Y(Wave 3)	X(Wave 3)
Unit 1:	y_{11}	x_{11}	y_{12}	x_{12}	y_{13}	x_{13}
Unit 2:	y_{21}	x_{21}	y_{22}	x_{22}	y_{23}	x_{23}
Unit 3:	y_{31}	x_{31}	y_{32}	x_{32}	y_{33}	x_{33}
.						
.						
Unit *n*:	y_{n1}	x_{n1}	y_{n2}	x_{n2}	y_{n3}	x_{n3}

- So if Y=Repression in a country and X=Democracy in a country, we would have Repression and Democracy scores for each country for different waves of observation in the same row of data.
- This, as we will also see in a few weeks, is SEM's preferred data structure. We can calculate correlations (covariances) between variables at different waves and these covariances are manipulated and "explained" with SEM analysis

Alternate “pooled” data structure: “Long” format

Unit	Time	Y	X
1	1	y_{11}	x_{11}
1	2	y_{12}	x_{12}
1	3	y_{13}	x_{13}
.	.	.	.
.	.	.	.
1	*t*	y_{1t}	x_{1t}
2	1	y_{21}	x_{21}
2	2	y_{22}	x_{22}
2	3	y_{23}	x_{23}
.	.	.	.
.	.	.	.
2	*t*	y_{2t}	x_{2t}
3	1	y_{31}	x_{31}
3	2	y_{32}	x_{32}
3	3	y_{33}	x_{33}
.	.	.	.
.	.	.	.
3	*t*	y_{3t}	x_{3t}
.	.	.	.
.	.	.	.
n	*t*	y_{nt}	x_{nt}
n	1	y_{nt}	x_{nt}
n	2	y_{nt}	x_{nt}
n	3	y_{nt}	x_{nt}
.	.	.	.
.	.	.	.
n	*t*	y_{nt}	x_{nt}

Each row represents a *unit at a given point in time*, and additional rows contains information on the variables from the same units at different points, followed by a series of rows that represents another unit’s information from time 1 to T, then rows that represents another unit’s information time 1 to time T, etc.

Continuing the Democracy-Repression example: the first row of data is Country 1, Wave 1 and that country’s score on Repression and Democracy; the second row of data is Country 1, Wave 2, and that country’s score on Repression and Democracy, followed by rows for all waves for Repression and Democracy for Country 1. Then country 2 follows in a series of rows, country 3, and so on, all the way to country *n*, time *t*.

- This is the general structure of "POOLED" CROSS SECTION-TIME SERIES DESIGNS
  - SOME HAVE LARGER T, SMALLER N (TSCS)
  - SOME HAVE LARGER N, SMALLER T (Panels)
- In either case, it gives an effective *N* of ***N*T*** in "balanced" designs, where all cases are observed at all points in time, or ΣT_i from 1 to N in "unbalanced" designs, where some cases are missing at some points in time
- Long Format easily can accommodate analyses with many panel waves, Wide Format too but becomes somewhat more cumbersome
- STATA can move easily from Long to Wide Form data (and back) with the **"reshape"** command
- Note the **"multilevel"** structure of Long Form Data: You have waves of observation "nested" within units, just like, e.g., individual students "nested" within classrooms. Wave of observation is "Level 1" and Unit is "Level 2" in this framework. We'll pick up on this idea and its analytic implications as the course progresses.

Multiwave Pooled Panel Regression

$$(1) \qquad Y_{it} = \alpha + \beta_1 X_{i1t} + \ldots \beta_k X_{ikt} + \beta_m Z_{im} + \varepsilon_{it}$$

- 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.

$$(1)\quad Y_{it} = \alpha + \beta_1 X_{i1t} + \ldots\ \beta_k X_{ikt} + \beta_m Z_{im} + \varepsilon_{it}$$

- 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
- What is not here at the moment?
  - Among other things, lagged Y, the "lagged endogenous variable." We will add that term later, but for the moment, and in many analyses within the econometric tradition, lagged Y is *not* an independent variable. Some of this is simply due to tradition and style, some due to substantive arguments we will consider
  - The "U" term representing unit-level unobserved heterogeneity/"stable unobservables" – we'll get (back) to this shortly
- All country-years are pooled together into one regression equation – there is not a separate model for wave 1, for wave 2, etc. (as we will see in the SEM tradition)

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 at the moment, 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 intrinsic 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.

$$(2) \qquad \mathrm{Y}_{\mathrm{it}} = \alpha + \beta_1 \mathrm{X}_{\mathrm{i1t}} + \ldots \beta_k \mathrm{X}_{\mathrm{ikt}} + \beta_m Z_{\mathrm{im}} + U_i + \varepsilon_{\mathrm{it}}$$

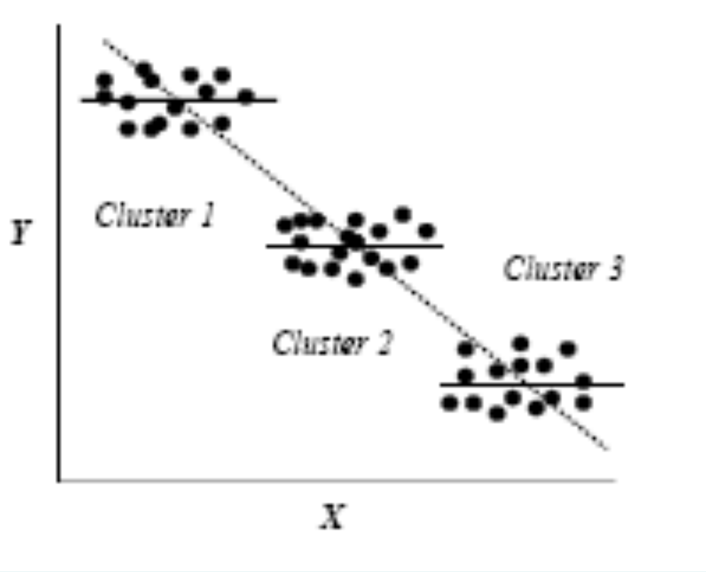
- With U_i representing all of the "***unobserved heterogeneity***," or the unobserved *stable* factors in case or unit *i*. (Unobserved "unstable" factors are still in the error term ε).
- The error term in this model is now composed of two parts: a unit-level effect that does not vary across time (U_i) and an idiosyncratic error term that varies across units and across time ($\varepsilon_{\mathrm{ij}}$). This composite error term ($\mathrm{U}_\mathrm{i} + \varepsilon_{\mathrm{ij}}$) 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 taking 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 + ε_{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 one 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!

$$(2) \qquad Y_{it} = \alpha + \beta_1 X_{i1t} + \ldots \beta_k X_{ikt} + \beta_2 Z_i + U_i + \varepsilon_{it}$$

- 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 will discuss later in the context of hierarchical or multilevel panel models

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."

$$(3) \qquad \mathrm{Y}_{\mathrm{it}} = \alpha + \beta_1 \mathrm{X}_{1\mathrm{it}} + \ldots \beta_k \mathrm{X}_{\mathrm{ikt}} + \beta_m Z_i + \mathrm{c}_1 D_1 + \mathrm{c}_2 D_2 + \mathrm{c}_3 D_3 + \ldots \mathrm{c}_{\mathrm{n\text{-}1}} D_{n-1} + \varepsilon_{\mathrm{it}}$$

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., **not** to be confused with "treatment variable" D from last session)!

- 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:

$$(4) \qquad \bar{Y}_{\mathrm{i}} = \alpha + \beta_1 \bar{X}_{1\mathrm{i}} + \beta_2 \bar{X}_{2\mathrm{i}} + \ldots \beta_k \bar{X}_{k\mathrm{i}} + \beta_m \bar{Z}_{m\mathrm{i}} + \bar{\mathrm{U}}_{\mathrm{i}} + \bar{\varepsilon}_{\mathrm{i}}$$

- 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:

$$Y_{it} - \bar{Y}_{\mathrm{i}} = (\alpha - \alpha) + \beta_1 (X_{1it} - \bar{X}_{1\mathrm{i}}) + \ldots \beta_k (X_{ikt} - \bar{X}_{\mathrm{ik}}) + \beta_m (Z_{im} - \bar{Z}_{\mathrm{im}}) + (\mathrm{U}_{\mathrm{i}} - \bar{\mathrm{U}}_{\mathrm{i}}) + (\varepsilon_{\mathrm{it}} - \bar{\varepsilon}_{i})$$

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

$$(5) \qquad Y_{it} - \bar{Y}_{\mathrm{i}} = \beta_1 (X_{1it} - \bar{X}_{1\mathrm{i}}) + \ldots \beta_k (X_{ikt} - \bar{X}_{\mathrm{ik}}) + \varepsilon^{*}_{it}$$

$$(5) \qquad Y_{it} - \bar{Y}_{\mathrm{i}} = \beta_1 (X_{1it} - \bar{X}_{1\mathrm{i}}) + \ldots \beta_k (X_{ikt} - \bar{X}_{\mathrm{ik}}) + \varepsilon_{it}^{*}$$

- 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 (or FD) 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 called an F* test:

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

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 automatically makes this adjustment in its 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 FD 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 Σ(XB) from the FE (***NOT*** LSDV) estimation, and correlate this with actual Y.
- Between R^2: Generate the AVERAGE of Σ(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 Σ(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:

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

  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:

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

(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

$$(6) \quad \mathrm{Y}_{\mathrm{it}} = \alpha + \beta_1 \mathrm{X}_{1\mathrm{it}} + \ldots \beta_k \mathrm{X}_{\mathrm{ikt}} + \beta_m Z_{mi} + t_1 T_1 + t_2 T_2 + \ldots t_{t-1} T_{t-1} + U_i + \varepsilon_{\mathrm{it}}$$

- 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)
- This is the equivalent of the time-varying intercepts in FD models, which capture the effect of time on the "control" group which always has "0" on the treatment variable
- And following this logic, the "t" effects must be assumed to be the same as what the "treatment" group's "t" would have been in the absence of treatment, as we discussed in the last session!

- 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. Other options exist (see text and readings)

Extensions: Time-Varying Effects of Covariates

- In the FE (and FD) models, time-invariant variables drop out, but *only* so long as their effects are assumed to be constant over time. (See slides 50-53 from last session's PPT). In FE, this assumption can be relaxed by including an *interaction* between Z and the time dummies (or time trend):

$$(7)\ \mathrm{Y}_{\mathrm{it}} = \alpha + \beta_1 \mathrm{X}_{1\mathrm{it}} + \ldots \beta_k \mathrm{X}_{\mathrm{ikt}} + \beta_m Z_{mi} + t_1 T_1 + \beta_{mt1} Z_m * T_1 + t_2 T_2 + \ldots t_{t-1} T_{t-1} + U_i + \varepsilon_{\mathrm{it}}$$

- 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

The Multi-wave First Difference (FD) Model

- Alternative way to estimate the unobserved heterogeneity model and take account of the U_i in multi-wave models is very simple: stack the data and estimate a first difference model, this time with multiple "first differences": the difference between waves 2 and 1, 3 and 2, 4 and 3, etc., all the way to T and T-1:

$$\text{(8a)} \quad Y_{it} - Y_{\text{it-1}} = (\alpha - \alpha) + \beta_1(X_{1it} - X_{1it-1}) + \ldots\beta_k(X_{kit} - X_{\text{kit-1}}) + \beta_m(\mathrm{Z}_{\mathrm{i}} - \bar{Z}_{\mathrm{i}}) + (\bar{\mathrm{U}}_{\mathrm{i}} - \bar{\mathrm{U}}_{\mathrm{i}}) + (\varepsilon_{\mathrm{i2}} \text{-} \varepsilon_{\mathrm{i1}})$$

$$\text{(8b)} \quad \Delta Y_{\mathrm{i}} = \beta_1 \Delta X_{1it,t-1} + \ldots\beta_k \Delta X_{kit,t-1} + \Delta\varepsilon_{it,t\text{-}1}$$

- The U_i is "swept out" of the equation through the differencing process (along with the intercept and all the other stable Z variables)
- An expanded FD model would include time-dummies to estimate the effect of given time period shocks on changes in all units, over and above any changes in the Xs:

$$\text{(9a)} \quad Y_{it} - Y_{\text{it-1}} = (\alpha - \alpha) + \beta_1(X_{1it} - X_{1it-1}) + \ldots\beta_k(X_{kit} - X_{\text{kit-1}}) + (t_t T_t - t_{t-1} T_{t-1}) + (\bar{\mathrm{U}}_{\mathrm{i}} - \bar{\mathrm{U}}_{\mathrm{i}}) + (\varepsilon_{\mathrm{i2}} \text{-} \varepsilon_{\mathrm{i1}})$$

$$\text{(9b)} \quad \Delta Y_{\mathrm{i}} = \beta_1 \Delta X_{1it,t-1} + \ldots\beta_k \Delta X_{kit,t-1} + \Delta t_{it,t\text{-}1} + \Delta\varepsilon_{it,t\text{-}1}$$

- In this formulation, the Δt is exactly the same as Δα in equation (6) from last week's two-wave FD model exposition – it is the "time effect" for all units, whether or not they changed on X (or "D" in last session's terminology)

Notes on FD

- FD similar in spirit to FE: both are difference equations, with FD subtracting out previous wave's raw value, and FE subtracting out the unit-specific mean. Both are thus "within" estimators
- The FD error terms are autocorrelated, since, for example, the error $\Delta\varepsilon$ from the time *t* and time *t-1* difference, and the time *t-1* and *t-2* difference both contain ε_{it-1}. So multi-wave FD models must be estimated with "Generalized Least Squares" (GLS) or other methods than can take this into account
- FD models can be expanded to include different lags and durations of the effects of "treatments". For example, the treatment might affect *all* subsequent values of Y, or the effect might decay in subsequent waves after an initial increase, and other possibilities. See the textbook and especially Allison (1994) for seminal treatments of this subject. (Also possible in FE).
- FD models can be easily expanded to include differential effects of *positive* changes in X versus *negative* changes in X (e.g., do *increases* in democracy have the same magnitude (negative) change in repression as *decreases* in democracy have on (positive) change? See Allison (2019) and our do file example.

FE or FD?

- With two-wave data, FE and FD give you *exactly the same results* – with one proviso, that the time dummy or two-wave FE model is estimated. A "naïve" FE model without the time effect will not pick up the changes in units that never change on X, what we described as the "control group" in the two-wave QE FD scenario. Including the time effects in the two-way FE model controls for this, and yields identical results for all X (and D) as FD with two wave data
- Generally, the FE is used more frequently in multi-wave panel analysis.
  - FE is more efficient than FD, in that it uses *more* information than does the FD when one is constructing the difference score. FD uses *only* the prior wave's value, while FE uses a composite average of *all* values of X or Y for its demeaning process. (One could ask why we should use only the lag value of X or Y in the FD model. Why not X minus lag(2) of X? or lag(3) of X?).
- FD used in some situations, however, notably when lagged Y is in the model (as we will see with "dynamic panel models" later in the course)

- FD model is more closely linked to “treatment effects” and the counterfactual model of causality, though that is changing.
- What is the value of the counterfactual outcome for a given case in the FE framework? It is “the average outcome from all the other time points for that unit that were observed”. The value of Y at a given point in time is compared to the mean outcome for Y for that case (“de-meaned” Y), so we are predicting mean-deviated Y as the “effect” of a series of X treatments
- Imai and Son (2019) propose a “weighted” FE model which uses the outcome at all **opposing treatment status waves** as the counterfactual for a given unit at a given time (instead of the total within-unit average outcome). This seems very sensible! (Module only available in R as of 2019)

Strengths and Weaknesses of the FE/FD 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 and especially FD 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/FD cannot correct for biases due to "endogenous selection", or selection into treatment due to *time-varying* unobservables that affect Y_{t-1}