# Section 13 Models for Pooled and Panel Data

## *Data definitions*

- **Pooled data** occur when we have a "time series of cross sections," but the observations in each cross section do not necessarily refer to the same unit.
    - HGL is ambiguous about this and sometimes use pooled to refer to panel data
- **Panel data** refers to samples of the *same* cross-sectional units observed at multiple points in time. A panel-data observation has two dimensions: $x_{it}$, where $i$ runs from 1 to $N$ and denotes the cross-sectional unit and $t$ runs from 1 to $T$ and denotes the time of the observation.
    - A **balanced panel** has every observation from 1 to $N$ observable in every period 1 to $T$.
    - An **unbalanced panel** has missing data.
    - Panel data commands in Stata start with **xt**, as in xtreg. Be careful about models and default assumptions in these commands.
- **Asymptotic properties**
    - We need to be careful about large-sample properties of these estimators: are we talking about $N \to \infty$, about $T \to \infty$, or both?

## *Regression with pooled cross sections*

- The crucial question with pooled cross sections from different time periods is "Does the same model apply in each time period?"
    - Has inflation changed the real values of some variables, requiring adjustment?
    - Was the business cycle at different phases in different periods?
    - Were there changes in technology or regulation that would cause behavior to be different?
    - Are there other factors that might cause coefficients in one period to differ from those in others?
- This is a special case of the Assumption #0 question: Do all observations come from the same model?
- **Time dummy variables**
    - A very general way of modeling (and testing for) differences in intercept terms or slope coefficients between periods is the use of time dummies.
    - Including time dummies (for all but one, omitted date in the sample to avoid the dummy-variable trap) alone allows the intercept to have a different value in each period.

        - The estimated intercept term in the model with time dummies is the estimated intercept in the period with the omitted dummy.
        - The estimated coefficient on an included time dummy corresponding to a particular period is an estimate of the difference between the intercept in that period and the intercept in the omitted period.
        - A joint test of whether all the dummies' coefficients are zero tests the hypothesis that the intercept does not vary at all over periods.
        - The simple test of whether a particular dummy's coefficient is zero tests the hypothesis that the intercept in that dummy's period does not differ from that of the omitted period.
    - Including interactions between time dummies and another variable *Z* allows the coefficient on (effect of) *Z* to vary across periods.
        - As before, the estimated coefficient on non-interacted *Z* is the estimated effect in the period for which the dummy is omitted.
        - The estimated coefficient on the interaction between *Z* and the dummy for period *t* is the estimated difference between the effect of *Z* in period *t* and the effect in the omitted period.
        - The joint test of the interaction terms tests the hypothesis that the coefficients (effects) of *Z* are the same in all periods.
        - The simple test of the interaction term for the period *t* dummy tests whether the effect of *Z* in period *t* differs from the effect in the omitted period.
- **Using aggregate variables that vary only over time (not across units)**
    - Suppose that we think that the reason for variation in either the intercept or a slope coefficient across periods is due to changes in one particular variable (the aggregate unemployment rate, for example).
    - In this case, we can include that variable (for intercept effects) and perhaps interactions of that variable with some regressor *Z* (to capture effects on unemployment on the marginal effect of *Z*).
    - Interpretation of these coefficients is standard for continuous interaction variables.
- **Limitations on variables that vary only over time**
    - If we include time dummies, we cannot include any other variables that vary only over time.
        - Any variable that varies only over time can be expressed as a linear function of the dummies.
            - If there are two periods with unemployment = 4 in the first period and 6 in the second, then $U = 4 + 2D_2$ , where $D_2$ is a dummy equal to one in the second period. Thus, including $U$, $D_2$, and a constant will result in perfect multicollinearity.

        - Same thing happens with more periods and/or more variables like *U* that vary only over time (and not across units).
    - If there are *T* time periods represented in the data, there can be at most $T - 1$ only-time-varying variables in the regression (assuming no dummies).
        - Again, there can be only *T* distinct "observations" for any such variable, so just as *N* must be at least *k* + 1 in a standard regression, we can only identify the effects of $T - 1$ such variables. Otherwise we have perfect multicollinearity.
        - We must also be careful about degrees of freedom here, because although we may have a large *N*, if we have only *T* = 2, we don't really have much information about the effect of the one time-only-varying observation whose effect we can estimate.

## *Pooled estimation with panel data*

- Simplest method is just to estimate by OLS with a sample of *NT* observations, not recognizing panel structure of data
    - Standard OLS would assume homoskedasticity and no correlation between unit *i*'s observations in different periods (or between different units in the same period)
    - **Clustered standard errors** are standard errors that are robust to correlation between error terms of same unit and heteroskedasticity over time:
        - $\text{cov}\left(e_{it}, e_{is}\right) = \psi_{ts},$ $\text{var}\left(e_{it}\right) = \psi_{tt}.$
        - Option vce(cluster) in Stata

## *Unit (entity) fixed effects*

- How might model vary across units?
    - All coefficients might be different: $y_{it} = \beta_{1i} + \beta_{2i}x_{2it} + \beta_{3i}x_{3it} + e_{it}$
        - Implies need to estimate separate regressions for each *i*
        - Impractical if *T* is small.
    - Constant terms might be different but slopes the same: $y_{it} = \beta_{1i} + \beta_2 x_{2it} + \beta_3 x_{3it} + e_{it}$
        - This is **unit fixed-effects model**
        - Can be estimated in two ways: LSDV or with de-meaned data
            - Former is impractical with large *N*
        - Intercept (dummy variable) estimates only converge asymptotically as $T \to \infty$.
- **Least-squares with unit dummy variables**

  - Introduce a dummy variable for each unit (individual) in sample:
    $y_{it} = \sum_{j=1}^{N} \beta_{1j} D_{ji} + \beta_{2i} x_{2it} + \beta_{3i} x_{3it} + e_{it}$
  - We now have $N + K - 1$ regressors with $NT$ observations. If $T$ is moderately large (> 2 or 3) then we can get reliable estimates, but note computational difficulty: moment matrix is $N + K - 1 \times N + K - 1$, which with $N > 100$ is probably computationally difficult.
    - Note that even if computing power is sufficient to blast through the problem, inverting very large matrices can be subject to more rounding error than simpler problems.
  - Can test the block of unit dummies to see if individual fixed effect is statistically significant.

- **Fixed-effects estimator using de-meaned data**
  - $y_{it} = \beta_{1i} + \beta_2 x_{2it} + \beta_3 x_{3it} + e_{it}$
  - Averaging across $T$ time periods for each unit $i$:
    $\frac{1}{T}\sum_{t=1}^{T} y_{it} = \beta_{1i} + \beta_2 \frac{1}{T}\sum_{i=1}^{T} x_{2it} + \beta_3 \frac{1}{T}\sum_{t=1}^{T} x_{3it} + \frac{1}{T}\sum_{t=1}^{T} e_{it}$
    $\bar{y}_i = \beta_{1i} + \beta_2 \bar{x}_{2i} + \beta_3 \bar{x}_{3i} + \bar{e}_i$
  - Subtracting the means from the original equation yields $y_{it} - \bar{y}_i = \beta_2 \left( x_{2it} - \bar{x}_{2i} \right) + \beta_3 \left( x_{3it} - \bar{x}_{3i} \right) + \left( e_{it} - \bar{e}_i \right)$ or $\tilde{y}_{it} = \beta_2 \tilde{x}_{2it} + \beta_3 \tilde{x}_{3it} + \tilde{e}_{it}$ where the ~ indicates deviation from the unit mean
  - This is the "within-unit estimator"
    - It uses only variation over time within each unit to identify the coefficients
    - If one *x* varies mostly across units rather than over time, there is not going to be much information in the sample to allow it to be identified by fixed-effects estimation.
    - If a variable varies *only* across units (e.g., ethnicity or sex), then its effects cannot be identified at all in a fixed-effects model
      - All ~ values will be zero because each observation equals the unit mean.
      - This also happens in LSDV because the *x* in question will be perfectly collinear with the unit dummies.
  - The constant term is gone because both $\tilde{y}_{it}$ and $\tilde{x}_{kit}$ have zero means.
    - We can estimate the mean of the individual unit constant terms from "between unit estimator" $\bar{y}_i = \bar{\beta}_1 + \beta_2 \bar{x}_{2i} + \beta_3 \bar{x}_{3i} + \left( \bar{e}_i + \beta_{1i} - \bar{\beta}_1 \right)$
    - We can calculate the estimated intercept for any individual unit as $\hat{\beta}_{1i} = \bar{y}_i - \hat{\beta}_2 \bar{x}_{2i} - \hat{\beta}_3 \bar{x}_{3i}$ under the assumption that $E\left( \bar{e}_i \right) = 0$

- Degrees of freedom
    - Although there are *NT* observations on the ~ variables, only $N(T-1)$ of them are independent.
    - Should use $\hat{\sigma}^2 = SSE / \left(NT - N - K + 1\right)$ to reflect this: An FE estimator will correct this but if you de-mean yourself and use OLS it will not.
    - Note that using LSDV will do this automatically because there will be *N* – 1 additional coefficients being estimated.
- An alternative to de-meaning the data is to subtract one time period from all the others.

## *Time fixed effects*

- If there are characteristics (especially unobserved ones) that are common to all units but vary across time, then we can use **time fixed effects**, which are just like the time dummies that we discussed in the pooling section.
- The model is then $Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}$. We omit the constant term if all *T* dummies are used to avoid collinearity; alternatively, we can omit the dummy for one time period.
- The methods of estimation are identical to the unit fixed-effects model.
    - We can, equivalently
        - Estimate the model with time dummies, or
        - Estimate with *y* and *x* expressed as deviations from time-period means across units.
- Any variable that varies only across time, and not across units, will be collinear with the dummy variables (or zero when de-meaned) and its effect cannot be estimated.
- We can also combine both unit and time fixed effects.
    - Either LSDV with both unit and time dummies, or
    - Demeaning the data both with respect to time and with respect to units.
        - To do this, we calculate $\tilde{y}_{it} = \left(y_{it} - \bar{y}_i\right) - \frac{1}{N}\sum_{j=1}^{N}\left(y_{jt} - \bar{y}_j\right) = \left(y_{it} - \bar{y}_i\right) - \left(\bar{y}_t - \bar{\bar{y}}\right)$, where $\bar{\bar{y}}$ is the overall mean across both units and time, and regress it on a similarly transformed *x*.
        - This is sometimes called the "differences-in-differences" estimator because it excludes the effects of changes that are strictly over time (taken out with time dummies or demeaning) *and* the effects of changes that are strictly across units (taken out with unit dummies or demeaning). This leaves only differences across units in how the variables change over time to estimate β.

## *Random-effects models*

- The fixed-effects model thinks of $\beta_{1i}$ as a fixed set of constants that differ across *i*.
- The random-effects model thinks of $\beta_{1i}$ as a random variable (with mean $\bar{\beta}_1$) that has one value for each *i* drawn from a given probability distribution.
    - $$\begin{aligned} \beta_{1i} &= \bar{\beta}_1 + u_i \\ E\left(u_i\right) &= 0 \\ \operatorname{cov}\left(u_i, u_j\right) &= 0,\ i \neq j \\ \operatorname{var}\left(u_i\right) &= \sigma_u^2 \end{aligned}$$
    - $$\begin{aligned} y_{it} &= \left(\bar{\beta}_1 + u_i\right) + \beta_2 x_{2it} + \beta_3 x_{3it} + e_{it} \\ &= \bar{\beta}_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \left(u_i + e_{it}\right) \\ &= \bar{\beta}_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + v_{it} \end{aligned}$$
    - This leads to a particular pattern of correlation among the error terms $v_{it}$
        - Error terms of observations corresponding to the same *i* will be correlated because they have $u_i$ in common
            - $\operatorname{corr}\left(v_{it} v_{is}\right) = \dfrac{\operatorname{cov}\left(v_{it}, v_{is}\right)}{\sqrt{\operatorname{var}\left(v_{it}\right)\operatorname{var}\left(v_{is}\right)}} = \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} = \rho$
        - Error terms of observations with different *i* will be uncorrelated (by assumption)
    - We can estimate $\rho$ by looking at the correlation of error terms within units and use the estimated $\hat{\rho}$ to do feasible GLS: this is the **random-effects estimator**
- Assumptions of the random-effects estimator
    - $E\left(v_{it}\right) = 0$
    - $\operatorname{var}\left(v_{it}\right) = \sigma_e^2 + \sigma_u^2$
    - $\operatorname{cov}\left(v_{it}, v_{is}\right) = \sigma_u^2,\ t \neq s$
    - $\operatorname{cov}\left(v_{it}, v_{js}\right) = 0,\ i \neq j$
    - $\operatorname{cov}\left(e_{ij}, x_{kij}\right) = 0,\ k = 2, 3, \ldots, K$
    - $\operatorname{cov}\left(u_i, x_{kit}\right) = 0,\ k = 1, 2, \ldots, K$
- Testing for presence of random effects vs. OLS assumption of independent errors:
    - Is $\sigma_u^2 = 0$?
        - If so, then there are no correlations because the *u* error term is degenerate
        - A positive value of $u_i$ is a positive expected value of $v_{it}$ for that *i*.
            - We can estimate with $\hat{u}_i = \dfrac{1}{T}\sum_{t=1}^{T} \hat{v}_{it}$.

        - Then estimate $\sigma_u^2 = \frac{1}{N}\sum_{i=1}^{N} \hat{u}_i$ .
    - One-tailed LM test for $H_0 : \sigma_u^2 = 0$ against $H_1 : \sigma_u^2 > 0$ in large samples is

$$LM = \sqrt{\frac{NT}{2(T-1)}}\left[\frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\hat{v}_{it}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{v}_{it}^2} - 1\right] \sim N(0,1)$$

    - If we reject $\sigma_u^2 = 0$ , then OLS is inefficient and we should use random effects (if assumptions are satisfied)
- Estimation in presence of random effects
    - OLS is unbiased, but inefficient. Can use clustered standard errors to correct bias in standard errors.
    - Efficient FGLS estimator:
        - First use OLS residuals to estimate $\sigma_u^2$ and $\sigma_e^2$ .
        - Calculate $\alpha = 1 - \frac{\sigma_e}{\sqrt{T\sigma_u^2 + \sigma_e^2}}$
        - Transform model by "quasi-de-meaning"

$$y_{it}^* \equiv y_{it} - \alpha\bar{y}_i$$
$$x_{1it}^* \equiv 1 - \alpha$$
$$x_{kit}^* \equiv x_{kit} - \alpha\bar{x}_{ki},\ k = 1,2,\ldots,K$$
$$v_{it}^* \equiv v_{ij} - \alpha\bar{v}_i$$

        - With this transformation, the *v** error term is homoskedastic and not autocorrelated, so we can apply OLS to the transformed model to get efficient estimators
        - Note that the random-effects estimator = fixed-effects estimator if α = 1, which is the limit as $T \to \bullet$ .
- Critical assumption of random-effects model is that the unit-specific error term (representing missing variables that are constant within each unit over time) are independent of the included regressors.
    - This is often not sensible.

## *Random effects or fixed effects?*

- Fixed-effects models have the advantage of not requiring cov(*x*, *u*) = 0, which is often difficult to justify.
    - Fixed-effects models are fully efficient as *N* gets large even if the true model is random effects.

- However, standard fixed-effects models cannot identify the effects of any variables that vary only across units (and has difficulty identifying effects if most of the meaningful variation is across units).
- Can do a Hausman test to examine whether the random-effects model is OK. (It is a nested sub-model of the fixed-effects model.)
    - The Hausman test is rejected if
        - The estimates are sufficiently different, ***and***
        - The fixed-effects estimators are sufficiently precise.
    - Common decision rule: Use random-effects unless the Hausman test rejects it.
- **Hausman-Taylor estimator** (xthtaylor)
    - Applies IV methods to random-effects model to overcome the correlation between the regressors and the error term so that we can use RE estimation to estimate the effects of variables that do not vary across time within individuals
    - $y_{it} = \beta_1 + \beta_2 x_{it,exog} + \beta_3 x_{it,endog} + \beta_4 w_{i,exog} + \beta_5 w_{i,endog} + u_i + e_{it}$
    - The *w* variables do not vary over time for any individual, the *x* variables do
    - The number of $x_{it,exog}$ must be ≥ the number of $w_{i,endog}$ in order for identification to be possible.
    - This estimation uses the demeaned $x_{it,endog}$ as instruments for the quasi-demeaned values and the means across time of $x_{it,exog}$ as instruments for the $w_{i,endog}$

## *SUR and panel data*

- Panel estimation by random effects is a lot like SUR because we are taking account of the same kind of correlation across observations.
- With small *N* and large *T*, it sometimes makes sense to treat the model as SUR and we can test whether the coefficients are the same across *i* as cross-equation coefficient restrictions.

## *Class demonstration?*

- Dataset: S&W's Seatbelts.dta
    - Show dataset
    - Define as panel
        - xtset fips year
    - Discuss missing values problem
    - Note two state identifiers, one alpha and one numeric
- Following S&W's Empirical Exercise E10.2
    - Generate lnincome variable
    - Discuss expected results of regressing fatalityrate on sb_usage speed65 speed70 drinkage21 ba08 lnincome age
- OLS regression

    - sb_usage has "wrong" sign
        - Authors argue that this is omitted variable bias and might be corrected partially by including state fixed effects.
    - Other effects are plausible
    - Send to outreg2 using fatal , word ctitle(OLS)
- FE regression
    - Now sb_usage has the expected sign
    - Other variables decline in coefficient magnitude
    - With and without the "robust" option, which gives clustered standard errors in this case.
    - ctitle(FE)
- Adding time dummies
    - What would time dummies control for?
        - Changes over time in such things as air bags, other auto or highway safety features
    - xi: xtreg … i.year , fe
    - Note loss in significance of all variables except age and drinkage, which become stronger.
    - This is "differences in differences" estimator and has very high bar for significance.
    - ctitle(FE & time)
- RE regression
    - Results are similar to other methods
    - Note requirement that the error term be uncorrelated with regressors
        - FE output has measure of the correlation between the fe dummies and the regressors.