Section 8 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.
- **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.

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.

Before and after estimators with panel data

- Simplest case of panel data is $T = 2$, where we can think of them as before and after some change of interest (perhaps a policy change).
- Advantage of panel data is that we have multiple observations with the same unobserved characteristics.
- Suppose that we have a balanced panel with $T = 2$ for $Y_{it} = \beta_0 + \beta_1 X_{it} + Z_i\gamma + u_{it}$, where Z_i is a matrix of variables (and γ the corresponding coefficients) that do not vary over time.
  - These are time-invariant "individual characteristics."
  - In the case of individuals, they might be ability, education, demographic characteristics, etc.—anything that we can assume doesn't change between the two dates of our sample.
  - In the case of states or countries, they might include laws or cultural factors that can plausibly be assumed constant over time (at least over the two time periods in our sample).
- For each i, if the two dates are $t = 1$ and $t = 2$, then $Y_{i,1} = \beta_0 + \beta_1 X_{i,1} + Z_i\gamma + u_{i,1}$, and $Y_{i,2} = \beta_0 + \beta_1 X_{i,2} + Z_i\gamma + u_{i,2}$.
  - Notice that the $Z_i\gamma$ part of the equation is the same in both periods. This is the essence of the before and after estimator.
- Taking the difference over time yields

$$\begin{aligned} Y_{i,2} - Y_{i,1} &= (\beta_0 - \beta_0) + \beta_1 (X_{i,2} - X_{i,1}) + (Z_i\gamma - Z_i\gamma) + (u_{i,2} - u_{i,1}) \\ &= \beta_1 (X_{i,2} - X_{i,1}) + (u_{i,2} - u_{i,1}). \end{aligned}$$

- Because the Z variables dropped out, it doesn't matter whether or not we can observe them. Thus, the before and after estimator is a great way of dealing with unobservable variables such as ability, effort, etc. (as long as they don't vary over time for any given unit in the panel).
- This equation can be estimated by OLS without a constant (noconstant option in Stata) as long as the differenced error term has the required properties.

- Note that it is critically important that X vary over time within units (and in a different way across units), otherwise the regressor in the differenced regression is zero (or constant).
- This estimator is nice when $T = 2$, but what if $T > 2$?
    - We can generalize it as the "fixed-effects" estimator.

Unit (entity) fixed effects

With $T > 2$, we could do $T - 1$ differences across pairs of time periods, allowing $n(T - 1)$ observations in the differenced sample (and $n(T - 1) - k$ degrees of freedom because there is no constant term). Alternatively, we can get a similar (identical if $T = 2$) regression in two other ways.

- **Regression with unit dummy variables**.
    - Let $D_i = 1$ for all observations on unit i and 0 otherwise, for $i = 2, 3, \ldots, n$. There are $n - 1$ such dummies.
    - We can run the **unit fixed-effects** regression $Y_{it} = \beta_0 + \alpha_i D_i + \beta_1 X_{1it} + \beta_2 X_{2it} + \ldots + \beta_k X_{kit} + u_{it}$ (generalizing S&W to more than one X).
    - Note that although we have nT observations in this regression, we end up with $nT - (n - 1) - k - 1 = n(T - 1) - k$ degrees of freedom, just as in the differenced case.
    - One can show that these two regressions are formally equivalent: the estimators for β are the same and have the same distributions, standard errors, etc.
    - Any X that has no variation across time within unit will be a linear function of the dummies, so we have perfect multicollinearity. We can't identify the effects of such non-time-varying variables in a fixed-effects model.
- **De-meaned regression**
    - Another equivalent way of estimating this model is to subtract the unit-mean from each observation.
        - Let $\bar{X}_i = \frac{1}{n}\sum_{i=1}^{T} X_{it}$ and $\bar{Y}_i = \frac{1}{n}\sum_{i=1}^{T} Y_{it}$.
        - Let $\tilde{X}_{it} = X_{it} - \bar{X}_i$ and $\tilde{Y}_{it} = Y_{it} - \bar{Y}_i$.
        - However, we really don't have nT independent observations because $\sum_{i=1}^{T} \tilde{X}_{it} = 0$, so $\tilde{X}_{iT} = -\sum_{i=1}^{T-1} \tilde{X}_{it}$ (and the same for Y and u).
        - In order to correct for this problem, we can either drop one time unit to eliminate the redundant observations or we can adjust the degrees of freedom to correct for this.
    - Most statistical packages actually do fixed-effects regression using the de-meaning procedure because it takes less time to calculate the means and the tilde

variables and invert a matrix of order $k + 1$ than to invert a matrix of order $k + n$, which would be necessary to estimate the model with $n - 1$ unit dummy variables. (Option fe in Stata xtreg, which is *not* the default)
    - This estimator is sometimes called the **within-unit estimator** because it estimates the coefficients strictly based on variation (over time) within cross-sectional unit.
        - Corresponding to this is a **between-unit estimator** that is the regression of $\bar{Y}_i$ on $\bar{X}_i$ for the n observations of the sample. (Option be in Stata xtreg)
    - Again, any variable that doesn't vary over time within each unit will have zero values from the deviation from mean, so regression breaks down and we can't identify the effect.
- It is worth thinking about where the most meaningful variation in your sample occurs.
    - Fixed-effect regression uses ***only*** changes over time within units in calculating the relationship among the variables.
    - If the meaningful variation in your sample is mostly ***between units*** (such as panel regressions on a sample of colleges, where the differences between colleges are much more important than the differences within colleges over time), then fixed-effect regression is unlikely to be effective.
        - If the variation is ***all*** between units, then we have perfect multicollinearity and we can't estimate the effects at all.
        - If ***most*** of the variation is between units, then we will have high (but not perfect) multicollinearity and our estimates will be very imprecise.
    - Because of this, fixed-effects regression sets a very high bar: if your effects are significant and meaningful in fixed effects you can probably attach considerable confidence to them.
- Is the fixed-effects model identical to the first-difference model?
    - Not if $T > 2$.
    - Although the data series span the same space, the assumptions made about the error terms are different.
    - If there is high correlation between $u_{i,t}$ and $u_{i,t-1}$, then the first-difference estimator is often better because the differencing eliminates this high correlation in a way that subtracting the mean does not.
    - If there is no strong correlation between adjacent (in time) observations, then the fixed-effects estimator is often better.
- The fixed-effects estimator is a straightforward application of OLS, and has the usual properties of the OLS estimator.
    - Stock and Watson list the assumptions of fixed-effects estimation in the box on page 365:
        - Conditional expectation of u conditional on X and α is zero.
        - IID draws.

        - Finite fourth moments of X and u.
        - No perfect multicollinearity.
        - Errors are uncorrelated within units across time.
    - Relaxation of the last assumption is possible by using **clustered standard errors** that correct for this pattern of autocorrelation (, vce(robust) in xtreg, assuming that covariance is positive only within units i)
        - S&W's appendix 10.2 discusses the formulas underlying clustered standard errors.
        - This variance estimator allows for nonzero covariance terms among the observations within a cross-sectional unit when calculating the error variance.
- **Testing significance of fixed effects**
    - This is just a standard F test of the hypothesis that all the coefficients of the dummy variables are zero.
    - Although Stata probably doesn't estimate with dummies, it does provide a test statistic for this hypothesis.
- **Asymptotic properties of the unit fixed-effects estimator**
    - What do we mean by asymptotic? $n \to \infty$? $T \to \infty$? Both?
        - Traditionally, econometricians think of T as fixed and $n \to \infty$ for asymptotic results in panel data models.
        - With this assumption, the estimates of β are consistent, but the estimates of α_i are unbiased but not consistent, because the number of observations used to estimate each α_i does not get large as $n \to \infty$ and thus the variance of the α estimators does not go to zero.
    - If both n and T get large, then estimates of β and α_i are both consistent.
- **Unbalanced panels**
    - All of these methods work with in principle unbalanced panels, although the statistical package may balk.

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.
    - With two units, we can take differences across units.
    - With $n > 2$, we can, equivalently
        - Estimate the model with time dummies, or

        - Estimate with Y and X expressed as deviations from time means.
- 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 $\tilde{Y}_{it} = \left(Y_{it} - \overline{Y}_i\right) - \left(\overline{Y}_t - \overline{\overline{Y}}\right)$, where $\overline{\overline{Y}}$ 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

Sometimes rather than having a different fixed constant term for each unit, we want to think of each unit having a common error term drawn randomly from some distribution. In other words, the α_i terms are thought of as random variables drawn, for each i, from some common distribution rather than constants.

- The model is $Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \ldots + \beta_k X_{kit} + \left(\alpha_i + u_{it}\right)$, where we have grouped α and u together as a composite error term.
    - This is sometimes called an **error-components** model because the error term has two components:
        - One that is the same across time within units, and
        - An "idiosyncratic" error term for each unit/period.
- In order for OLS to be consistent for the random-effects model, we must be able to assume that α is uncorrelated with X: The unobserved, time-invariant characteristics of units that influence the dependent variable must be uncorrelated with the measured variables whose effects we want to estimate.
- The composite error terms (v) of observations within the same group are correlated: if α and u are independent of one another, then $\text{cov}\left(v_{it}, v_{is}\right) = \dfrac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2}$.
    - The random-effects estimator is a feasible GLS estimator that estimates this covariance based on correlation between same-unit residuals, then calculates an estimator that is BLUE conditional on this calculated covariance matrix.

- Computationally, this involves "quasi-de-meaning" the data by calculating $\tilde{Y}_{it} = Y_{it} - \lambda \bar{Y}_i$, where $\lambda = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2} \right)^{\frac{1}{2}}$. As $T \to \infty$, $\lambda \to 1$ and the random-effects estimator becomes equivalent to the fixed-effects model.

- **Random effects or fixed effects?**
  - Fixed-effects models have the advantage of not requiring cov(X, α) = 0, which is often difficult to justify.
  - However, 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 appropriate. (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.
    - Use random-effects unless the Hausman test rejects it.

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 endogeneity 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.