Fixed Effects Models for Panel Data

December 1, 2014

Notation

Use the same setup as before, with the linear model

$$\mathbf{Y_{it}} = \mathbf{X_{it}}\beta + \mathbf{c_i} + \epsilon_{\mathbf{it}} \tag{1}$$

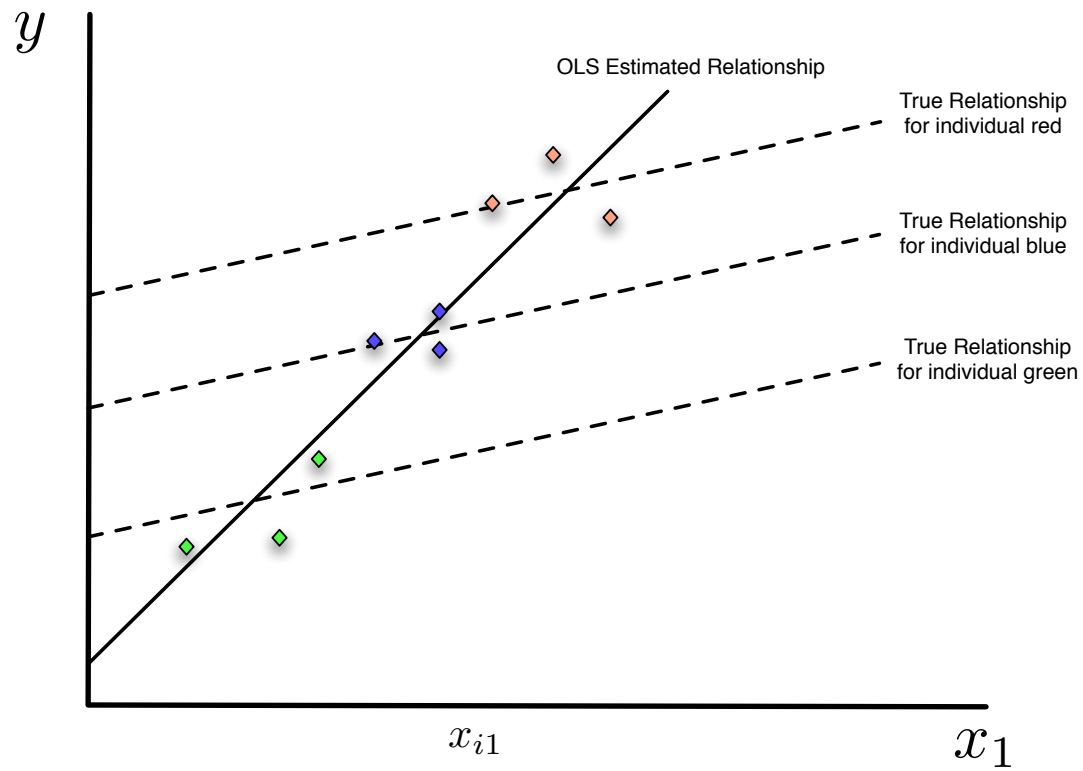
where X_{it} is a $1 \times K + 1$ vector of independent variables. Here we make our "usual assumptions":

Assumption 1: $E[\epsilon_{it}|X_{i1}, \ldots, X_{iT}, c_i] = 0$

Assumption 2: $E[\epsilon_i \epsilon_i'] = \sigma^2 I_T$

What about $E(\mathbf{X}'\mathbf{c})$?

In the fixed effects model, we do not have to make assumptions about whether unobserved heterogeneity is correlated with our independent variables. So it can handle this case:



Method 1: The Dummy Variable Estimator

Add an $N \times 1$ dimensional time invariant vector, called Z to the regression model, where Z looks like this for observation $N = 1$

$$Z_1 = \begin{bmatrix} 1 \dots 0 \end{bmatrix} \tag{2}$$

Using this approach, we can write the estimating equation as

$$\mathbf{Y_{it}} = \mathbf{X_{it}}\beta + \mathbf{Z_i c} + \epsilon_{\mathbf{it}} \tag{3}$$

Just an OLS Estimator

$$\underset{\mathbf{c},\mathbf{b}}{min} S(\mathbf{b}) = (\mathbf{Y} - \mathbf{Xb} - \mathbf{Zc})' (\mathbf{Y} - \mathbf{Xb} - \mathbf{Zc}) \tag{4}$$

To see this consider $X = []$, and only the fixed effects dummies are included. In that case,

$$\hat{c}_i = \sum_{t=1}^{T} \frac{Y_{it}}{T} \tag{5}$$

Having variance $\frac{\sigma^2}{T}$. Note this does not approach zero as $N \to \infty$

Lots of c_i's to estimate and not consistent

The variance of ϵ, σ^2 is

$$\frac{1}{NT-(K+1)-N}\hat{\epsilon_{\mathbf{dv}}}'\hat{\epsilon_{\mathbf{dv}}} \tag{6}$$

where

$$\hat{\epsilon}_{\mathbf{dv}} = (\mathbf{Y} - \mathbf{Xb} - \mathbf{Zc}) \tag{7}$$

Besides the consistency problem, this estimator requires the identification of $(N-1)+K+1$ parameters (and inverting an $(N-1)\times K+1)$ square matrix.

Method 2: The Demeaning Estimator

To overcome the consistency problems with the dummy variable estimator, most statistical packages employ the "Demeaning Estimator". This estimator does not fit a constant for each cross section unit N but importantly, it *does not* merely put the unobserved heterogeneity effect c_i into the error term.
To proceed, define

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it} \quad \textit{and} \quad \bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it} \tag{8}$$

and writing the deviations from the means for X and Y as

$$\ddot{Y}_{it} = Y_{it} - \bar{Y}_i \quad \textit{and} \quad \ddot{X}_{it} = X_{it} - \bar{X}_i \tag{9}$$

The Demeaning Estimator, cont.

We can recover the demeaning estimator (b_d) for β by OLS on the demeaned data as

$$\mathbf{b_d} = (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\ddot{\mathbf{Y}} \tag{10}$$

Sometimes also referred to as the "With-In Estimator".
The "Between" estimator uses only variation between cross section units. It is recovered by the regression $\bar{Y}_i = \bar{X}_i\beta + c_i + u_i$

Standard errors from the Simple OLS approach needs a little tweaking to be correct

While the regression estimates are unbiased, since $E(\ddot{\epsilon}|\mathbf{X}_i) = 0$, the standard variance covariance matrix from simple OLS is not correct. To see this, recall that

$$\begin{aligned} E[\ddot{\epsilon}_{it}] =& E[(\epsilon_{\mathbf{it}} - \bar{\epsilon}_{\mathbf{i}})^{\mathbf{2}}] = E[\epsilon_{it}^2 - 2\epsilon_{it}\bar{\epsilon}_i + \bar{\epsilon}_i{}^2] && (11)\\ =& \sigma^2 + \sigma^2/T - 2\sigma^2/T && (12)\\ =& \sigma^2(1 - 1/T) && (13) \end{aligned}$$

and

$$\begin{aligned} E[\ddot{\epsilon}_{it}\ddot{\epsilon}_{is}] =& E[(\epsilon_{it} - \bar{\epsilon}_i)(\epsilon_{is} - \bar{\epsilon}_i)] && (14)\\ =& 0 - \sigma^2/T - \sigma^2/T + \sigma^2/T && (15)\\ =& -\sigma^2/T && (16) \end{aligned}$$

Demeaning Estimator Parameter Variance Covariance Matrix

$$
\begin{aligned}
Var(\mathbf{b_d}|x) =& E[(b_d - b)(b_d - b)'|x] && (17)\\
=& E[(\ddot{X}'\ddot{X})^{-1}\ddot{X}'\ddot{Y} - \beta)(\ddot{X}'\ddot{X})^{-1}\ddot{X}'\ddot{Y} - \beta)'] && (18)\\
=& E[(\ddot{X}'\ddot{X})^{-1}\ddot{X}'(\ddot{X}\beta + \ddot{\epsilon}) - \beta)(\ddot{X}'\ddot{X})^{-1}\ddot{X}'(\ddot{X}\beta + \ddot{\epsilon}) - \beta)'] && (19)\\
=& \sigma_u^2 E[(\ddot{X}'\ddot{X})^{-1}] && (20)
\end{aligned}
$$

$$
\hat{\sigma_u^2} = \frac{(\ddot{Y} - \ddot{X}\beta)'(\ddot{Y} - \ddot{X}\beta)}{N(T-1) - (K+1)} \qquad (21)
$$

Method 3: First Differences

Recall the first differencing approach we discussed back in the introduction. Suppose that for each individual, we have a panel of two periods ($t = 1, 2$). Apply a differencing approach for each individual i to rid the model of c_i, since

$$\begin{aligned} \Delta y_i =& \beta (x_{i2} - x_{i1}) + (c_i - c_i) + (\epsilon_{i2} - \epsilon_{i1}) && (22) \\ =& \Delta x_i \beta + \Delta \epsilon_i && (23) \end{aligned}$$

Then, we have another way of estimating the model that

- Rids the model of the c_i
- But does not put them in the error term

The estimating equation becomes:

$$\Delta y = \Delta x \beta + \Delta \epsilon \tag{24}$$

Having estimates equal to

$$\mathbf{b_{fd}} = (\mathbf{\Delta x' \Delta x})^{-1} \mathbf{\Delta x' \Delta y} \tag{25}$$

These are all equivalent methods for $T = 2$- that is, you will recover the same β estimates.

3 ways to estimate, which one to use?

- If $T = 2$, these will all yield exactly equivalent results for the parameters and the variance/covariance matrices.
- Demeaning is preferred in most cases since it does not require the estimation of N constants

Consider the following small dataset on N=2 and T=3

Person	Year	Wage	Education	Experience	Training
1	2000	12	12	5	0
1	2001	15	12	6	1
1	2002	15	12	7	1
2	2000	25	16	0	0
2	2001	27	16	1	0
2	2002	30	16	0	1

For each of the 3 methods: Dummy Variable Estimator, Demeaning Estimator, and the First Differences Estimator construct the matrix of independent variables for the model.

Testing for Endogeneity

Steps:

- Partition your matrix of explanatory variable for each individual as $\mathbf{x_i} = [\mathbf{x_{i1}} \quad \mathbf{x_{i2}}]$. Note that $\mathbf{x}_{i1}$ is the subset of exogenous independent variables and x_{i2} is the the potentially endogenous explanatory variable having an instrument z_{i2}.
- Run the relevancy test regression

$$\Delta x_{i2t} = \Delta \mathbf{x}_{i1t}\mathbf{a} + \Delta \mathbf{z}_{i2t} b + \Delta u_{it}, t = 2, ..T \tag{26}$$

  and recover $\Delta \hat{u}_{it}$
- As in the endogeneity chapter,

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\beta + \Delta \hat{u}_{it}\delta + \psi \tag{27}$$

  And test for $H_0 : \delta = 0$. Rejecting H_0 is a strong signal of an endogeneity problem.

OLS versus RE

This test relies on our finding that

$$\lim_{\sigma_c^2 \to 0} \Omega = \frac{1}{\sigma_\epsilon^2} I_{NT \times NT} \tag{28}$$

Restricting $\sigma_c^2 = 0$ means the model can't fit the data as well, so the test stata does (the Breusch Pagan LM test) checks to see how much our predictive power is degraded due to the restriction. I am being intentionally vague since this is a Maximum Likelihood-based test (have not covered yet).

This tests

H_0: $\sigma_c^2 = 0 \Rightarrow$ Pooled OLS appropriate
H_1: $\sigma_c^2 \neq 0 \Rightarrow$ Random Effects Appropriate

Fixed Versus Random Effects

The critical difference between the random and fixed effects approaches is whether c_i is correlated with $\mathbf{x}_{it}$. If it is, then the random effects approach, which simply puts c_i in the error term will lead to biased estimates relative to the fixed effects estimator. As in the Chapter on endogeneity, we need to test to see how "different" the two estimated β vectors are and to do this, we use the Hausman test.

Hausman Test

$$H = (b_{RE} - b_{FE})' \left[V\hat{A}R(b_{RE}) - V\hat{A}R(b_{FE}) \right]^{-1} (b_{RE} - b_{FE}) \quad (29)$$

is distributed with χ^2_M, and M are the degrees of freedom in the model.

This tests
H_0: $E[\mathbf{x}'\mathbf{c} = 0] \Rightarrow \mathbf{E}[\mathbf{b^{re}}] = \mathbf{E}[\mathbf{b^{fe}}] = \beta \Rightarrow$ RE
H_1: $E[\mathbf{x}'\mathbf{c} \neq 0]$ so $\mathbf{E}[\mathbf{b^{re}}] \neq \mathbf{E}[\mathbf{b^{fe}}] = \beta \Rightarrow$ FE

When to use Pooled OLS, RE, FE

- Pooled OLS Estimator: When the unobserved heterogeneity is not present. All assembly line workers receive the same training, are more or less of the same background, and only workers of a certain skill level are retained. Therefore, productivity rates only vary randomly across people.
- Random Effects: Sequential measurement of astronomical phenomena. We observe a star day in a day out and make every known correction to the measurement of its position. That position may have measurement error and the error may contain a piece that varies day in and day out and the other piece may shift the error in a systematic way across observation.

When to use Pooled OLS, RE, FE

- Fixed Effects: Almost all examples from economics are a fixed effects story. Aid received by recipient countries probably depend on unobserved factors specific but invariant through time, that are also correlated with other recipient country factors. Unobserved factors on aid allocation by donors related to quality of aid implementation, and this is correlated with democratic institutions.