Econometrics II

Prof. Brendan Kline

Homework on Panel Data

Due on October 16

This homework asks you to consider some questions relating to panel data:

1. Consider the standard model

$$y_{it} = x_{it}\beta + c_i + u_{it} \tag{1}$$

for $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$. Suppose that all elements of x are time varying for at least some individuals (in particular, so that all estimators below are well-defined, in the sense that the full rank conditions are true). Note that $x_{it} \in \mathbb{R}^d$ where $d \geq 1$. In other words, x_{it} can be a *vector* of explanatory variables that affect y_{it}.

Let $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}$, and $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$.

The fixed effects transformation results in:

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + u_{it} - \bar{u}_i. \tag{2}$$

Another implication of the standard model results from averaging within individuals:

$$\bar{y}_i = \bar{x}_i\beta + c_i + \bar{u}_i \tag{3}$$

The *total* estimator is the pooled OLS estimator applied to equation 1. In other words, the *total* estimator is the linear regression of y_{it} on x_{it}. Denote the corresponding estimator $\hat{\beta}_{total}$. This is a "naive" estimator that makes no attempt to deal with the panel nature of the data. Essentially it just uses NT observations.

The *within* estimator is the pooled OLS estimator applied to equation 2. In other words, the *within* estimator is the linear regression of $y_{it}-\bar{y}_i$ on $x_{it}-\bar{x}_i$. Denote the corresponding estimator $\hat{\beta}_{within}$. This is the same as the *fixed effects* estimator.

The *between* estimator is the pooled OLS estimator applied to equation 3. Denote the corresponding estimator $\hat{\beta}_{between}$. In other words, the *between* estimator is the linear regression of $\bar{y}_i$ on $\bar{x}_i$.

(a) Prove that

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)'(x_{it}-\bar{x}_i)+T\sum_{i=1}^{N}\bar{x}_i'\bar{x}_i=\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}'x_{it}$$

(b) Prove that

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)'(y_{it}-\bar{y}_i)+T\sum_{i=1}^{N}\bar{x}_i'\bar{y}_i=\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}'y_{it}$$

(c) Prove that

$$\hat{\beta}_{total}=W(x)\hat{\beta}_{within}+(I-W(x))\hat{\beta}_{between},$$

where $W(x)$ is a matrix-valued function of the data $\{x_{it}\}_{i=1,t=1}^{N,T}$ (of the appropriate dimension) and I is the identity matrix (of the appropriate dimension). Provide an explicit expression for $W(x)$. Hint: write down the standard formula for pooled OLS corresponding to each of these estimators. The terms in parts 1a and 1b of the question should appear in those expressions. One approach is to start by rewriting the standard formula for $\hat{\beta}_{total}$ using parts 1a and 1b. Some of the terms in the rewritten formula for $\hat{\beta}_{total}$ also appear in the standard formulas for $\hat{\beta}_{within}$ and $\hat{\beta}_{between}$. So by rearranging the standard formulas for $\hat{\beta}_{within}$ and $\hat{\beta}_{between}$... This question shows there is an algebraic relationship between the three estimators.

2. Wooldridge Problem 10.3 (about FE and FD estimators with two periods of panel data). Do only part (a).

Hint: Use facts like the following: when $T = 2$, $x_{i1} - \bar{x}_i$ can be written in terms of the difference between x_{i1} and x_{i2}. These sorts of facts make it possible to start from the standard formula for $\hat{\beta}_{FE}$ and rewrite to get the standard formula for $\hat{\beta}_{FD}$.

3. Wooldridge Problem 10.4 (about program evaluation with two periods of panel data). Do only parts (a)-(c).

   Hint for (c): In a regression of y on binary x (and an intercept), with $y = \alpha + x\beta + u$, OLS basically finds the solutions to these equations: $E(y|x = 0) = \alpha$ and $E(y|x = 1) = \alpha + \beta$.

4. Based on Wooldridge Problem 10.10 (about murder rates). The `murder.dta` dataset is panel data on murder rates and factors potentially affecting murder rates. In particular, the research question is: do executions (of criminals convicted of murder) deter individuals from committing murder? In other words, the research question is: what is the effect of executions on the murder rate? There is a huge empirical literature debating this issue, and this question asks you to explore some related issues in the context of our study of panel data.

   The dataset concerns the various states in the United States (plus D.C.), and has data from 1987, 1990, and 1993. The dataset includes three variables relevant to this question: $mrdrte_{it}$ is the murder rate in state i during year t (the number of murders in state i during year t, per 100000 people); $exec_{it}$ in state i during year t is the total number of executions for the current and prior two years; and $unem_{it}$ is the current unemployment rate in state i during year t. This question requires using Stata.

   Use the `list` command to view the data. By inspecting the values of the `state` and `year` variables note, for example, that the state "AL" (Alabama) appears three times in the dataset, for the years 87, 90, and 93. Also note that the murder rate (`mrdrte`) in Alabama in 1987 was 9.3, in 1990 was 11.6, and in 1993 was also 11.6. Note that the `id` variable also indicates the state, as a number. So, `id` = 1 indicates Alabama, `id` = 2 indicates state "AK" (Alaska), and so forth.

An unobserved effects model explaining current murder rates in terms of the number of executions in the last three years and the current unemployment rate is

$$mrdrte_{it} = \alpha + \beta_1 exec_{it} + \beta_2 unem_{it} + c_i + u_{it}.$$

(a) First, estimate the model using ordinary least squares, not accounting for the panel nature of the data. That is, use the command `regress mrdrte exec unem`. What is the estimated effect of executions on the murder rate?

(b) Really, we want to estimate the model by fixed effects. In order to do so, we must tell Stata how to identify the unit of observation (the i index, in this case state), and how to identify the time period of observation (the t index, in this case year). This is done using the `xtset` command. Specifically, use the command `xtset id year`. The first variable (`id` in this case) indicates the unit of observation, and the second variable (`year` in this case) indicates the time period of observation. As always, you can use the command `help xtset` to learn more about `xtset`.

Now, we can use the `xtreg` command to estimate the model using fixed effects. Specifically, use the command `xtreg mrdrte exec unem, fe cluster(id)`. The `, fe` option specifies fixed effects estimation. The `, cluster(id)` option specifies computing clustered[1] standard errors, where the clusters are the states (see above for connection to the `id` variable). Therefore, the `, cluster(id)` option affects the standard errors but not the estimates of β. Verify that by using the command `xtreg mrdrte exec unem, fe`. Again, you can use the command `help xtreg` to learn more about `xtreg`. What is the estimated effect of executions on the murder rate?

(c) Fixed effects estimation can be carried out via the "dummy variable regression" described in section 10.5.3 of Wooldridge. To do that, we need to create dummy vari-

[1]Clustered standard errors will be one of the later topics covered in lecture. It is possible to answer these questions without any knowledge of clustered standard errors, simply because the question doesn't ask anything about the standard errors!

ables that indicate state. Specifically, use the command `tabulate id, generate(stdum)`. And use the command `list` again to see what happened. Notice that 51 new dummy variables were added to the dataset (50 states, plus D.C.) and that they indicate the state. So for example, `stdum1` = 1 exactly for observations of state 1 (i.e., `id` = 1). And `stdum2` = 1 exactly for observations of state 2 (i.e., `id` = 2). And so forth. Now do the "dummy variable regression" by using the command `regress mrdrte exec unem stdum1-stdum50`. What is the estimated effect of executions on the murder rate?

(d) Compare the three estimates of the effect of executions on the murder rate. Give an intuitive explanation for why the estimate from part 4a is different from the other two estimates, in terms of omitted variables bias (where the omitted variable is the c_i term.)

(e) Under what circumstances would $exec_{it}$ not be strictly exogenous (conditional on c_i)? Of course, to answer this, you will need to discuss the meaning of strict exogeneity in this application.

Be sure to submit your Stata log / Stata output! Indicate using pen/pencil which part of the log/output corresponds to each part of the question.