ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Selection On Unobservables Designs: Part 1, Fixed Effects and Random Effects Models[1]

Panel data models loosely qualify under the rubric of selection on unobservables designs because they assume that individual-specific time series variation is a valid source of variation for identifying causal effects (i.e. it is as good as randomly assigned).

Panel, or *longitudinal*, data sets consist of repeated observations for the same units, firms, individuals or other economic agents. Typically the observations are at different points in time. Let Y_{it} denote the outcome for unit i in period t, and X_{it} a vector of explanatory variables. The index i denotes the unit and runs from 1 to N, and the index t denotes time and runs from 1 to T. Typically T is relatively small (as small as two), and N is relatively large. As a result, when we try to approximate sampling distributions for estimators, we typically approximate them assuming that N goes to infinity, keeping T fixed.

Here we will mainly look at *balanced* panels, where T is the same for each unit. An *unbalanced* panel has potentially different numbers of observations for each unit. This may arise because of units dropping out of the sample; a practical example would be firms going out of business.

The core research design issue that panel data addresses is the possibility that individual units may differ in important, unobserved ways that affect their outcomes in a manner that is constant over time. From a statistical standpoint, however, the key issue with panel data is that Y_{it} and Y_{is} tend to be correlated even conditional on the covariates X_{it} and X_{is}.[2]

[1] These notes are partially derived from Guido Imbens' old ARE 213 notes. Some passages have been quoted directly.

[2] Even if you are unconcerned about whether your estimates have a causal interpretation, the statistical issue still exists. In the context of the model below, if you assume that X_{it} is scalar and let $\beta = \frac{Cov(X,Y)}{Var(X)}$, you will still need to account for the fact that Y_{it} and Y_{is} are correlated.

Selection on Observables: $Y_i(k) \perp D_i | X_i \quad \forall k \in \{0,1\}$

Selection on Unobservables: find Z_i w/ $Y_i(k) \perp Z_i \quad \forall k \quad \& \quad Cov(D_i, Z_i) \neq 0$

Types of Covariates: ① Endogenous – would appear on LHS of some eq that describes the DGP, should be determined simultaneously? "bad controls" – could add more bias
② Predetermined – can control for, and should
often conflicted ③ Exogenous – could control for but don't need to, should be unbiased

tutoring → grades
↳ also ↗ hours spent on class
might want to grades = $\beta_0 + \beta_1$·tutoring + β_2·hours

Consider these possibilities in a linear model setting:

$$Y_{it} = X'_{it}\beta + c_i + \varepsilon_{it}.$$

Statistically, the presence of c_i, the unobserved individual effect, creates a correlation between Y_{it} and Y_{is} even if ε_{it} is uncorrelated over time and units. If we assume that $E[\varepsilon_{it}|X_{i1}, ..., X_{iT}, c_i] = 0$, then we can condition on the c_i to estimate the effect of x on y. For the moment, however, we focus on the statistical issue.

The two main approaches to dealing with such issues are *fixed effects* and *random effects*. The labels are, unfortunately, rather deceptive. The key distinction is whether we model the correlation between the individual effects, c_i, and the covariates, X_{it}, or whether we assume that they are independent. Better labels would therefore be correlated and uncorrelated random effects, but the labels fixed and random effects are, by this point, fixed.

In both cases we assume that the vectors of individual outcomes are independent *across* individuals.

The first assumption we make for random effects (and also for fixed effects) is

is c_i corr w/ X_{it}?
yes → FE
no → RE to correct the SE
↳ Generalized least Squares estimator

Assumption 1 (Strict Exogeneity)

$$\mathbb{E}[\varepsilon_{it}|X_{i1}, \ldots, X_{iT}, c_i] = 0.$$

Second,

Assumption 2 (uncorrelated effects)

$$\mathbb{E}[c_i|X_{1i}, \ldots, X_{iT}] = 0.$$

Note that these two assumptions imply (somewhat unrealistically) that the composite error term, $c_i+\varepsilon_{it}$, is uncorrelated with the explanatory variables X_{it}. Thus an OLS regression

of Y_{it} on X_{it} will produce causal estimates, assuming that the X_{it} can be interpreted as treatments.

We will examine some of the ideas and concepts with data from the Card and Krueger (1994) minimum wage paper, which examines the effect of a minimum wage increase in New Jersey on fast-food employment, using Pennsylvania restaurants as a control group. We will examine regressions of fulltime employment on wages:

$$\text{emp}_{it} = \beta_0 + \beta_1 \cdot \text{wage}_{it} + v_{it}.$$

1 OLS

Given the strict exogeneity and uncorrelated effects assumptions, we can write

$$Y_{it} = X_{it}'\beta + v_{it},$$

with $v_{it} = c_i + \varepsilon_{it}$, the composite error term, uncorrelated with the covariates. Thus we can use OLS to estimate β:

$$\hat{\beta}_{ols} = (X'X)^{-1}(X'Y) = \left(\sum_{i,t} X_{it} \cdot X_{it}'\right)^{-1} \left(\sum_{i,t} X_{it} \cdot Y_{it}\right).$$

unbiased & consistent
but wrong SE
need clustered SE

Under our two assumptions, OLS gives us consistent estimates of β. However, the OLS standard errors will be incorrect because they assume independence across observations. In fact, the existence of c_i implies that observations in different time periods within a given cross-sectional unit will be positively correlated, even if the c_i terms are independent across units (as we assumed). We can write the OLS standard errors as:

$$V(\hat{\beta}) = \sigma_v^2 \left(\sum_{i,t} X_{it}X_{it}'\right)^{-1} = \sigma_v^2 \cdot \left(\sum_i X_i'X_i\right)^{-1},$$

where X_i is the $T \times K$ matrix with tth row equal to X_{it}' (i.e., X_i contains all of the X_{it} for a given cross-sectional unit i).

We can get robust (clustered) variances that account for the correlated structure of the data by first estimating the residuals as

$$\hat{v}_{it} = Y_{it} - X_{it}'\hat{\beta}.$$

The robust variance is then

$$V(\hat{\beta}) = \left(\sum_i X_i'X_i\right)^{-1} \cdot \left(\sum_i X_i'\hat{v}_i\hat{v}_i'X_i\right) \cdot \left(\sum_i X_i'X_i\right)^{-1},$$

where $\hat{v}_i$ is the T vector with tth element equal to $\hat{v}_{it}$.

The variance estimator above is what you get when you specify the ", cluster(unit)" option in Stata, where "unit" is the variable containing the cross-sectional identifier. We will discuss the properties of this variance estimator in further detail in a later lecture.

We apply the clustered variance estimator to the CK data. In these data, the restaurant is the cross-sectional unit i, and there are only two time periods (pre-minimum wage increase and post-minimum wage increase). We find

$$\begin{array}{rccc} \widehat{\text{emp}_{it}} & = & 12.4430 & + \quad 1.1103 \cdot \text{wage}_{it} \\ (OLS) & & (4.4951) & (0.9328) \\ (Clustered) & & (3.9921) & (0.8284) \end{array}$$

SEs

The clustered variance estimates are, in this case, slightly smaller than the OLS variance estimates, suggesting that correlation within restaurants over time is not seriously biasing the conventional standard errors.[3] This is due in part to the fact that there are only two time periods, so the potential bias in the standard errors is limited.[4]

[3] There are, however, other issues with the standard errors in this research design which we will discuss later.

[4] Particularly because the treatment is negatively correlated within units lying in New Jersey! Again, we will discuss these issues in more depth later.

2 Random Effects (GLS)

Recall that in conventional (i.e., non-panel) data sets with heteroskedastic residuals, we have two options for estimation. One option is to simply estimate the coefficients via OLS (which remains consistent) and use the Eicker-Huber-White "robust" standard errors to get the standard errors right. The other is to model the heteroskedasticity as a function of X and to then use weighted least squares (a special case of GLS) to estimate the coefficients and standard errors. Both procedures should produce consistent coefficient estimates and standard errors, but the latter is theoretically more efficient than the former (assuming you can model the heteroskedasticity correctly).

The situation with panel data is not different. As we saw above, under our two assumptions we can use OLS to estimate the coefficients and then correct the standard errors using the clustered standard errors. However, in principle we should be able to leverage the correlated structure of the data to produce a GLS estimator that gets the standard errors right and is more efficient than OLS. This, in essence, is what the random effects (RE) estimator is.

To exploit some of the random effects structure, define

$$\Omega = \mathbb{E}[v_i v_i'],$$

where $v_i = (v_{i1}, \ldots, v_{iT})'$, i.e. it is the vector of all residuals for unit i. In other words, Ω is the within-unit variance/covariance matrix of the residuals – it defines how residuals for the same unit across different time periods are correlated with each other.

The RE error structure consists of an assumption that implies that all residuals within unit i are equally correlated with each other:

Assumption 3 (Random Effects Error Structure)

$$\Omega = \sigma_\varepsilon^2 \cdot I_T + \sigma_c^2 \cdot \iota_T \cdot \iota_T' = \begin{pmatrix} \sigma_\epsilon^2 + \sigma_c^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_\epsilon^2 + \sigma_c^2 & \dots & \sigma_c^2 \\ \vdots & & \ddots & \\ \sigma_c^2 & \dots & & \sigma_\epsilon^2 + \sigma_c^2 \end{pmatrix}.$$

Here ι_T is a column vector of dimension T with all elements equal to one, i.e. $\iota_T \cdot \iota_T'$ is a $T \times T$ matrix full of ones. However, we continue to assume that the residuals across different units are uncorrelated with each other.[5]

We can exploit the error structure by estimating the variance/covariance matrix and then using weighted least squares:

$$\hat{\beta}_{RE} = \left(\sum_i X_i' \hat{\Omega}^{-1} X_i \right)^{-1} \left(\sum_i X_i' \hat{\Omega}^{-1} Y_i \right).$$

T×T T×k

Here X_i is the $T \times K$ matrix with tth row equal to X_{it}'. The consistency of this estimator, like that for the OLS estimator, does not depend on the random effects error structure (we are, after all, simply reweighting the data – we are still assuming that all variation in X is valid variation for estimating β).

The estimator for Ω is

$$\hat{\Omega} = \begin{pmatrix} \hat{\sigma}_\epsilon^2 + \hat{\sigma}_c^2 & \hat{\sigma}_c^2 & \dots & \hat{\sigma}_c^2 \\ \hat{\sigma}_c^2 & \hat{\sigma}_\epsilon^2 + \hat{\sigma}_c^2 & \dots & \hat{\sigma}_c^2 \\ \vdots & & \ddots & \\ \hat{\sigma}_c^2 & \dots & & \hat{\sigma}_\epsilon^2 + \hat{\sigma}_c^2 \end{pmatrix},$$

To get estimates for σ_ε^2 and σ_c^2, first estimate β by OLS, then calculate the residuals

$$\hat{v}_{it} = Y_{it} - X_{it}' \hat{\beta},$$

$\sigma_\epsilon^2 + \sigma_c^2 \simeq \text{var}(\hat{v}_{it})$

$\sigma_c^2 \simeq \text{cov}(\hat{v}_{it}, \hat{v}_{is})$ — seems like a strong assumption,

↳ could run RE and apply clustered SE

[5] Thus if we sorted observations first by i and then by t, the variance/covariance matrix for all of the data would look like a block diagonal matrix, with Ω on the diagonal.

Estimate the residuals' variance as

$$\hat{\sigma}_v^2 = \frac{1}{N \cdot T - K} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{v}_{it}^2,$$

Note that this is simply the mean of the sum of squared residuals for the entire data set (with a degrees of freedom adjustment applied) – nothing fancy here, despite the double-sum notation.

Then estimate the variance of the unobserved individual effect as

$$\hat{\sigma}_c^2 = \frac{1}{NT(T-1)/2 - K} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it} \cdot \hat{v}_{is},$$

This is just the mean of the product of $\hat{v}_{it}$ and $\hat{v}_{it+u}$ over the entire data set (with a degrees of freedom adjustment). The $NT(T-1)/2$ term in the denominator is divided by 2 because the last sum starts at $t+1$ which, on average, is halfway to T.

Finally, estimate

$$\hat{\sigma}_\varepsilon^2 = \hat{\sigma}_v^2 - \hat{\sigma}_c^2$$

(or zero if this is negative).

Applying this procedure to the Card and Krueger data we get $\hat{\sigma}_v^2 = 79.8707$, $\hat{\sigma}_c^2 = 43.4978$, and $\hat{\sigma}_\varepsilon^2 = 36.3729$.

If the model is correct, including our specification for Ω, then the variance for $\hat{\beta}_{RE}$ is

$$V(\hat{\beta}) = \left(\sum_i X_i' \hat{\Omega}^{-1} X_i \right)^{-1}.$$

If the specification for Ω is not correct, we can still apply the random effects estimator (it's still consistent under our two original assumptions, it's just not efficient anymore) and use the robust (clustered) variance estimator to get the standard errors right:

$$V(\hat{\beta}) = \left(\sum_i X_i' \hat{\Omega}^{-1} X_i \right)^{-1} \cdot \left(\sum_i X_i' \hat{\Omega}^{-1} \hat{v}_i \hat{v}_i' \hat{\Omega}^{-1} X_i \right) \cdot \left(\sum_i X_i' \hat{\Omega}^{-1} X_i \right)^{-1}.$$

Note that the clustered variance estimator relaxes the assumption that the off-diagonal elements of Ω are all the same. In other words, it allows for different correlations between different time periods, whereas the conventional RE standard errors assume that the correlation between different observations for the same unit is always the same, regardless of how far apart the time periods are.

Applying the random effects estimator to the Card and Krueger data gives us

$$\begin{array}{rccc} \widehat{\text{emp}}_{it} & = & 11.6952 & + & 1.2659 \cdot \text{wage}_{it} \\ (GLS) & & (3.8631) & & (0.7995) \\ (Clustered) & & (3.2317) & & (0.6701) \end{array}$$

Note that the random effects point estimates are still in the same ballpark as the OLS point estimates. If we believe our assumptions, then the similarity of the point estimates is not surprising – both estimators are consistent under these assumptions.

3 Feasible Generalized Least Squares (FGLS)

→ for making less assumptions about Ω $\quad \hat{\Omega}_{GLS} = \begin{pmatrix} \hat{\sigma}_1^2 + \hat{\sigma}_c^2 & \hat{\sigma}_c^2 \\ \hat{\sigma}_c^2 & \hat{\sigma}_2^2 + \hat{\sigma}_c^2 \end{pmatrix}$ assumptions about functional form of Ω

$\hat{\Omega}_{FGLS} = \frac{1}{N} \sum_i \hat{v}_i \hat{v}_i'$ average of Ω elements over N units

better if you have large N since Ω elements are N averages

In the conventional random effects model, we make fairly strong assumptions about the structure of the within-unit covariance matrix, Ω. An alternative form of Feasible Generalized Least Squares (FGLS) that is less restrictive does not rely on assuming the exact structure of the covariance matrix of the residuals. Instead, it estimates the structure of the covariance matrix using an estimator that is similar to the robust clustered variance estimator that we have been using for the clustered standard errors.

Again we start with OLS estimates which are consistent but not efficient in a wide range of settings. Estimate the residuals v_i as $\hat{v}_{it} = Y_{it} - X_{it}'\hat{\beta}_{ols}$. Then estimate the residual covariance matrix as

$$\underset{T \times T}{\hat{\Omega}} = \frac{1}{N} \sum_{i=1}^{N} \hat{v}_i \hat{v}_i'.$$

$$\underset{NT \times NT}{\Omega} = \begin{bmatrix} \Omega_1 & & 0 \\ & \Omega_2 & \\ 0 & & \Omega_N \end{bmatrix}$$

Not realistic, but often intra-cluster cov. dominates inter-cluster cov. so could be good estimate

If RE assumptions are correct, RE will be more efficient. If those assumptions are incorrect, then FGLS will be more efficient.

Note that $\hat{v}_i\hat{v}_i'$ forms an estimate of the $T \times T$ covariance matrix from the data for each cross-sectional unit. Unlike the conventional random effects estimator, it does not impose the constraint that all of the off-diagonal elements be equal.

Finally, estimate β as

$$\hat{\beta}_{FGLS} = \left(\sum_i X_i' \hat{\Omega}^{-1} X_i \right)^{-1} \left(\sum_i X_i' \hat{\Omega}^{-1} Y_i \right).$$

The advantage of the FGLS estimator relative to the RE estimator in the case with more than two periods is that it allows for a more flexible correlation structure. The disadvantage is that if the RE restrictions are (close to being) satisfied, then you may introduce a lot of extra noise by not exploiting them. This will be particularly true if N is not too large – then the quantity $\frac{1}{N}\sum_{i=1}^{N} \hat{v}_i\hat{v}_i'$ will not be a very accurate estimate of Ω. This is because, regardless of the size of T, each element in the $\hat{\Omega}$ matrix is estimated by only N observations. The RE restriction that all of the off-diagonal elements are equal allows us, in contrast, to estimate the off-diagonal elements using $NT(T-1)/2$ observations.

With only two periods FGLS gives us results identical to those for the RE estimator because there is only one unique off-diagonal element in the Ω matrix. Thus the RE structure does not restrict the full variance/covariance matrix of the residuals at all.

We can test for the presence of individual unobserved effects using the statistic

$$S = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it} \cdot \hat{v}_{is} \approx \sqrt{N} \cdot \hat{\sigma}_c^2 . \cdot T(T-1)/2$$

If $\sigma_c^2 = 0$, then $\hat{\sigma}_c^2$ will plim to 0, and S will be asymptotically normal:

$$S \to \mathcal{N}(0, V),$$

Note that we are effectively treating each of the i units as being independent of each other, and saying that the sum of the unique off-diagonal elements in $\hat{\Omega}$ for each unit have

some variance V. In other words, think of the random variable as:

$$S_i = \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it} \cdot \hat{v}_{is}.$$

Then $S = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} S_i$, and we know the CLT will apply since the S_i are distributed i.i.d. (we can even relax the identical assumption – it's the independence part that is important). All we need to do, then, is estimate $\text{Var}(S_i) = V$.

We can easily estimate V using method of moments as

$$\hat{V} = \frac{1}{N} \sum_{i=1}^{N} S_i^2 = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it} \cdot \hat{v}_{is} \right)^2.$$

Under the null hypothesis, $\hat{\sigma}_c^2$ will plim to 0; under the alternative hypothesis, $\hat{\sigma}_c^2$ will plim to something else. Thus our test consists of testing whether $S/\sqrt{\hat{V}}$ is distributed $\mathcal{N}(0,1)$.[6]

In the Card and Krueger data, $\hat{\sigma}_c = 43.4978$, so $S = 823.0176$, $V = 21,455$ and the test statistic is 5.6188, so we can reject the hypothesis that $\sigma_c^2 = 0$.

4 Fixed Effects (FE)

The OLS, RE, and FGLS estimators all maintain the uncorrelatedness assumption, i.e.

$$\mathbb{E}[c_i | X_{1i}, \ldots, X_{iT}] = 0.$$

From a research design perspective, then, these estimators are not that interesting because they do not allow us to relax the assumption that our variables of interest, X_{it}, are uncorrelated with the composite error term, $v_{it} = c_i + \varepsilon_{it}$, i.e. the X's that we're interested in still need to be "as good as randomly assigned" with respect to the errors.

[6] Note that the denominator here is estimated using the second moment rather than the second centered moment. Even so, $S/\sqrt{\hat{V}}$ will go to infinity as N gets large under the alternate hypothesis, because S is proportional to $\sqrt{N} \cdot \hat{\sigma}_c^2$ while $\hat{V}$ just converges to a fixed quantity.

The Fixed Effects model (FE) allows us to relax this uncorrelatedness assumption. We now only assume strict exogeneity:

$$\mathbb{E}[\varepsilon_{it}|X_{i1},\ldots,X_{iT},c_i]=0.$$

In other words, we only need assume that ε_{it} is mean-independent of X_i after conditioning on the individual effects c_i. Any part of the composite error that is time-invariant will get folded into c_i, so we can relax the unconfoundedness assumption for any component in the error term that varies across individuals but not over time. For statistical inference, we actually make a strong assumption (that we will relax later):

Random Effects error structure assumption:

$$\mathbb{E}[\varepsilon_i\varepsilon_i'|X_i,c_i]=\sigma^2\cdot I_T$$

The idea behind fixed effects is that we want to either estimate the c_i parameters (so that we can control for them) or just get rid of them altogether (so that we don't have to worry about them). We first consider the traditional fixed effects estimator, which consists of simply throwing in a whole mess of dummy variables, one for each unit i.[7] Formally, we implement this by adding an N-dimensional vector of covariates, R_{it}, with its jth element for unit i in period t equal to:

N-1 to avoid dummy variable trap

$$R_{it,j}=\mathbf{1}(i=j)$$

In other words, the first element of R equals unity if a given observation corresponds to unit 1 and zero otherwise, the second element of R equals unity if a given observation corresponds to unit 2 and zero otherwise, and the Nth element of R equals unity if a given observation corresponds to unit N and zero otherwise. One dummy for each cross-sectional unit.

[7] Of course, to avoid perfect colinearity, you will have to exclude one of the dummy variables when running the regression.

We can then estimate the FE model using the following linear regression:

$$Y_{it} = X_{it}'\beta + R_i'c + \varepsilon_{it}$$

where c is now an $N \times 1$ column vector containing all N of the c_i terms. Our estimate of β and c will now be unbiased (given our assumptions), but our estimates of c are not consistent because we do not get more observations with which to estimate each coefficient c_i as we increase N. If this is not intuitive, consider a special case in which X_{it} does not exist. In that case, $\hat{c}_i = \sum_t Y_{it}/T = \overline{Y}_i$, and the precision of $\overline{Y}_i$ does not increase as T stays fixed and N increases.

4.1 Within Estimator = Fixed Effects Estimator

In practice, it can be hard to estimate the FE model if N is very large. If N were 10,000, for example, you would effectively be asking the computer to invert a greater than $10,000 \times 10,000$ matrix when you included R as matrix of regressors. Fortunately, it turns out that there is a simple demeaning transformation that we can apply to the data that gives us estimates of β that are numerically identical to those produced by the FE estimator. This estimator is known as the *within estimator* because it identifies β using within-individual variation. Because it generates the same estimates as the FE estimator, people sometimes use the terms FE and within estimator interchangeably.

Define the unit-specific averages for unit i as

$$\overline{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}, \qquad \text{and } \overline{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$$

Then define the deviations from the unit-specific means as

$$\ddot{Y}_{it} = Y_{it} - \overline{Y}_i, \qquad \text{and } \ddot{X}_{it} = X_{it} - \overline{X}_i,$$

The within estimator is based on running the regression:

$$\ddot{Y}_{it} = \ddot{X}'_{it}\beta + \ddot{\varepsilon}_{it}$$

Note that we no longer have to invert the mega-matrix in order to run this regression. The key here is that the c_i terms disappear because $\ddot{c}_i = c_i - \overline{c}_i = 0$. We are thus left needing to only satisfy the orthogonality condition:

$$\mathbb{E}[\ddot{\varepsilon}_{it}|\ddot{X}_{it}] = 0$$

This holds true under the strict exogeneity assumption that we began with (you can use iterated expectations to condition on c_i if you want to show this).

Intuitively, it should be clear that the within estimator is equivalent to the FE estimator if you consider doing partitioned regression. If you regress each variable in X on all of the individual specific dummy variables, R, then the coefficients on the dummies will be equal to the individual specific means (recall that regressing a variable on a column of ones estimates the mean of that variable; regressing a variable on an indicator variable for unit i estimates the mean of that variable for unit i). Thus, after we partial out R from X, we will have $\widetilde{X} = X - R\overline{X}$, where $\overline{X}$ is a $N \times K$ matrix in which the ith row consists of $\overline{X}'_i$. Thus $\widetilde{X}_{it} = X_{it} - \overline{X}_i$, which is identical to the within transformation.

We can also write the within estimator in matrix form by defining the $T \times T$ matrix:

$$A = I_T - \iota_T\iota'_T/T = \begin{bmatrix} 1-\frac{1}{T} & -\frac{1}{T} & \dots & -\frac{1}{T} \\ -\frac{1}{T} & 1-\frac{1}{T} & \dots & \dots \\ \vdots & \vdots & \ddots & \ddots \\ -\frac{1}{T} & \vdots & \ddots & 1-\frac{1}{T} \end{bmatrix}$$

Then

$$\ddot{X}_i = AX_i.$$

One notable property of A is that it is idempotent, so technically you could recover the within estimates by only applying the transformation to the X_i (you wouldn't have to do it to the Y_i).[8] If you wanted to apply the transformation to the entire $NT \times K$ matrix of data, X, then it would be

$$\ddot{X} = \mathbf{A}X$$

where $\mathbf{A}$ is a block-diagonal $NT \times NT$ matrix containing A as its diagonal elements.

We should note that the OLS standard errors will be incorrect when running the within estimator because $\ddot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$ will be correlated across different observations within the same unit. To get the correct standard errors when using the within estimator, multiply the OLS standard errors by $\sqrt{T/(T-1)}$.

4.2 Differencing Estimators

Another class of estimators that can get rid of the c_i terms are differencing estimators. The most common differencing estimator is the first differences estimator. First define:

$$\Delta Y_{it} = Y_{it} - Y_{it-1}, \quad \Delta X_{it} = X_{it} - X_{it-1}, \quad \Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-1}$$

Then we can run the regression

$$\Delta Y_{it} = \Delta X_{it}'\beta + \Delta\varepsilon_{it}$$

using data from time periods 2, ..., T. Note that the c_i terms drop out because they are constant within any cross-sectional unit over time. OLS on this regression will produce consistent estimates, though the standard errors will need to be adjusted for serial correlation (one easy way would be to use the robust clustered variance estimator).

[8] Note that $AA = (I_T - \iota_T\iota_T'/T)(I_T - \iota_T\iota_T'/T) = I_T - 2\iota_T\iota_T'/T + \iota_T\iota_T'\iota_T\iota_T'/T^2 = I_T - \iota_T\iota_T'/T = A$

When $T = 2$, $\hat{\beta}_{FE}$ and $\hat{\beta}_{\Delta}$ are numerically identical. You will be asked to show this in a problem set exercise.

More generally, we can define a differencing estimator over s periods as:

$$\Delta_s Y_{it} = Y_{it} - Y_{it-s}, \quad \Delta_s X_{it} = X_{it} - X_{it-s}, \quad \Delta_s \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-s}$$

Then run the regression

$$\Delta_s Y_{it} = \Delta_s X_{it}'\beta + \Delta_s \varepsilon_{it}$$

using data from time periods $s + 1$, ..., T. When we choose $s = T - 1$ – the largest possible value for s – we refer to the resulting estimator as the "long differences" estimator (it's the longest differencing estimator we can apply to the data).

4.3 Fixed Effects vs. Random Effects

We know that the FE estimator is equivalent to the within estimator – it estimates β using only within-individual variation.[9] This sort of qualifies as a selection on unobservables design in that we think that within-individual variation is a source of "good" variation in X, while between-individual variation is not a source of "good" variation in X. We get rid of the between-individual variation in X by including fixed effects or applying the within transformation.

The complement of the within estimator is, of course, the between estimator, $\hat{\beta}_B$ – it estimates β using only between individual variation. It can be easily implemented by running the regression:

$$\overline{Y}_i = \overline{X}_i'\beta + \overline{\varepsilon}_i \qquad \hat{\beta}_B = \text{between}$$

[9] In practice we often include a dummy variable for each time period so that we remove aggregate time trends as well.

It turns out that the RE estimator is a weighted average of the within estimator (i.e., the FE estimator) and the between estimator. First define the within-individual and between-individual sums of squares:

$$S_{xx}^w = \sum_i \sum_t (X_{it} - \overline{X}_i)(X_{it} - \overline{X}_i)'$$

sum of variation of individuals from their mean.

$$S_{xx}^b = \sum_i T(\overline{X}_i - \overline{X})(\overline{X}_i - \overline{X})'$$

# of periods

variation of individual from total individual average.

FE is a weighted avg of all the differencing estimators (not shown here), so in terms of efficiency (I think) it lives somewhere between the first-diff estimator and the last diff estimator.

Then the RE estimator equals

$$\hat{\beta}_{RE} = \hat{F}^w \hat{\beta}_{FE} + (I - \hat{F}^w)\hat{\beta}_B$$

less biased — more biased

where

$$\hat{F}^w = [S_{xx}^w + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_c^2} S_{xx}^b]^{-1} S_{xx}^w$$

$\sigma_c \nearrow \Rightarrow \hat{\beta}_{RE} \to \hat{\beta}_{FE}$

If X contains a single regressor and $\sigma_c^2 = 0$, then $\hat{F}^w$ is simply the proportion of total variation in X that is within-individual variation, and each estimator receives a weight equal to its proportion of the total variation in X. If $\sigma_c^2 > 0$, then $\hat{F}^w$ increases, i.e. the within estimator begins to receive more weight at the expense of the between estimator. The intuition here is that, if $\sigma_c^2 = 0$, then $\overline{v}_i$ is an average of several independent variables, so the idiosyncratic shocks will tend to cancel each other out. If $\sigma_c^2 > 0$, however, then $\overline{v}_i$ contains an individual-specific shock (c_i) that affects all observations for individual i. This individual-specific shock will not get cancelled out when taking the mean over different observations for the same individual, so the between-individual data will have more noise than it would if $\sigma_c^2 = 0$. Hence $\hat{\beta}_B$ has a higher variance, and the RE estimator puts less weight on $\hat{\beta}_B$ than it would if $\sigma_c^2 = 0$.

It should now be obvious why random effects is more efficient than fixed effects – RE uses both within-individual and between-individual variation for estimating β while FE uses only

within-individual variation for estimating β. If c_i is not correlated with X_i, then there is no reason to throw out all that between-individual variation.[10] On the other hand, it is also clear why RE and OLS are biased if the uncorrelated effects assumption is dropped, while FE remains unbiased. RE and OLS are averages of the within and between estimators. If $\mathbb{E}[c_i|X_i] \neq 0$, then the between estimator will be biased, and RE and OLS will be biased.

4.4 Measurement Error

Measurement error often becomes more problematic when working with panel data. Consider the "classical measurement error" scenario in which the true variable, x_i^*, is measured with error. The observed variable is $x_i = x_i^* + u_i$, where u_i has zero mean and is uncorrelated with x_i^* or the error term ε_i. In this model, it is straightforward to show that $\hat{\beta}_{OLS}$ converges to $\beta \frac{\sigma^2_{x^*}}{\sigma^2_{x^*}+\sigma^2_u}$. In other words, the degree of attenuation bias is a function of the signal-to-noise ratio.

Now consider the panel data estimators in the context of a panel with $T = 2$. In this panel, the FE/within estimator and the first differences estimator are numerically identical (technically the long differences estimator is also identical, since the longest possible lag is one time period). Consider a one-regressor case with classical measurement error.

$$\text{plim}(\hat{\beta}_{FE}) = \text{plim}(\hat{\beta}_{\Delta}) = \frac{\text{Cov}(\Delta x, \Delta y)}{\text{Var}(\Delta x)} = \frac{\text{Cov}(\Delta x^* + \Delta u, \Delta x^* \beta + \Delta \varepsilon)}{\text{Var}(\Delta x^* + \Delta u)}$$

Substituting in $\Delta x^* = x^*_{it} - x^*_{it-1}$ and $\Delta u = u_{it} - u_{it-1}$ and simplifying the expression above gives

$$\text{plim}(\hat{\beta}_{\Delta}) = \beta \frac{\sigma^2_x(1-\rho_x)}{\sigma^2_x(1-\rho_x) + \sigma^2_u(1-\rho_u)}$$

[10] OLS can also be written as a weighted mean of the within and between estimators. In the case of OLS, the formula is identical to the formula for $\hat{\beta}_{RE}$ except that it does not contain any of the σ^2 terms. This is not surprising – if we assume $\sigma^2_c = 0$, then the variance-covariance matrix in the RE model looks exactly like the standard OLS covariance matrix. Accordingly, assuming $\sigma^2_c = 0$ in $\hat{F}^w$ gives us the OLS weights.

where

$$\sigma_x^2 = \text{Var}(x_{it}^*) \quad \text{and} \quad \sigma_u^2 = \text{Var}(u_{it})$$

$$\rho_x = \frac{\text{Cov}(x_{it}^*, x_{it-1}^*)}{\text{Var}(x_{it}^*)} \quad \text{and} \quad \rho_u = \frac{\text{Cov}(u_{it}, u_{it-1})}{\text{Var}(u_{it})}$$

Suppose that we estimate β using only cross-sectional variation; in particular, suppose that we use the between estimator, $\hat{\beta}_B$. In that case, doing a similar derivation to the one above, we will find that:

$$\text{plim}(\hat{\beta}_B) = \beta \frac{\sigma_x^2(1+\rho_x)}{\sigma_x^2(1+\rho_x) + \sigma_u^2(1+\rho_u)}$$

When is the attenuation bias from using the within-individual variation in the panel data (i.e., the FE/first differences estimator) more severe than the attenuation bias from using the cross-sectional variation in the panel data (i.e., the between estimator)? In other words, when is

$\rho_x \nearrow \Rightarrow$ LHS $\uparrow$, RHS $\downarrow$

$\rho_u \nearrow \Rightarrow$ LHS $\downarrow$, RHS $\uparrow$

$$\frac{\sigma_x^2(1+\rho_x)}{\sigma_x^2(1+\rho_x) + \sigma_u^2(1+\rho_u)} > \frac{\sigma_x^2(1-\rho_x)}{\sigma_x^2(1-\rho_x) + \sigma_u^2(1-\rho_u)}?$$

It's possible to solve this by cranking through a ton of algebra, but it's easier to simply note that when $\rho_x = \rho_u$, $(= 0)$ then the two sides of the inequality are equal. But increasing ρ_x unambiguously increases the left side (the top rises faster than the bottom) and decreases the right side (the top falls faster than the bottom). And decreasing ρ_u unambiguously increases the left side and decreases the right side. So the inequality above is satisfied if and only if:

$$\rho_x > \rho_u$$

Thus we see that attenuation bias gets worse from using FE/first differences (in comparison to cross-sectional variation) when the inter-period correlation in x_{it}^* is greater than the inter-period correlation in u_{it}. This condition is likely to hold if u_{it} truly is random noise.

Intuitively, if x_{it}^* is highly correlated across time periods within a given individual, while u_{it} is relatively uncorrelated across time periods, then removing the individual-specific mean $\overline{x}_i^*$ removes a lot of variation from the "signal", but removing $\overline{u}_i$ does not remove much "noise." Hence the signal-to-noise ratio often gets worse when using FE or differencing estimators (relative to RE or OLS, both of which leverage some degree of between-individual variation).

When $T > 2$, FE and differencing estimators continue to have problems with attenuation bias, but the degree of bias now differs because the two estimators are no longer numerically identical. In general, if you are willing to assume that the correlation between x_{it}^* and x_{is}^* is higher when s and t are closer, then first differences should have more attenuation bias than fixed effects, and long differences should have less attenuation bias than fixed effects. A sensible strategy for exploring whether you have measurement error issues might then be:

(1) First estimate $\hat{\beta}_{RE}$ and $\hat{\beta}_{FE}$. If the two are very close, then stop (you don't appear to have any issues).

(2) If $|\hat{\beta}_{FE}| < |\hat{\beta}_{RE}|$ (which is often the case), estimate β using first differences ($\hat{\beta}_{\Delta FD}$) and long differences ($\hat{\beta}_{\Delta LD}$).

(3) If $|\hat{\beta}_{\Delta FD}| < |\hat{\beta}_{FE}| < |\hat{\beta}_{\Delta LD}|$, then you may have a measurement error problem (in which case you would probably prefer $\hat{\beta}_{\Delta LD}$). If FE, first differences, and long differences are all similar, it is more likely that FE is less than RE because the uncorrelated effects assumption on c_i is false.

Of course, there is no guarantee that $|\hat{\beta}_{\Delta FD}| < |\hat{\beta}_{FE}| < |\hat{\beta}_{\Delta LD}|$ implies a measurement error problem. You could alternatively argue that the long differences estimator is more vulnerable to omitted variables bias from individual-specific trends than the FE and first differences estimators.

4.5 Application: Deschenes and Greenstone (2007)

Deschenes and Greenstone (2007) is an application that highlights both the potential advantages and weaknesses of the fixed effects approach. They use panel data on agricultural profits from 1978 to 2002 (collected at 5 year intervals) to estimate the effects of climate change on agriculture. Early work on this question examined the cross-sectional relationship between climate (e.g., temperature and rainfall) and the value of agricultural land. These estimates of the effect of climate on agricultural profits are valid (and, in fact, clearly superior to fixed effects estimates) if there is no correlation between average climate across different counties and unobserved characteristics that may affect agricultural profits. For example, states in the Midwest have different climates than states in the Northeast, and both have different climates than California. Cross-sectional estimates will be valid if we believe that other factors that affect agricultural profits and differ across states are uncorrelated with the average climate of a state.

We may not want to make the assumption that cross-sectional variation in average climate is uncorrelated with other factors affecting agricultural. Deschenes and Greenstone apply a fixed effects estimator to their panel data to remove the cross-sectional variation. Specifically, they include county fixed effects in their model, so that all of the variation occurs at the within-county level over time.[11] Since weather is pretty close to "as good as randomly assigned" on a year-to-year basis, this fixed effects model is likely to recover the causal effect of short-term climate fluctuations on agricultural profits.

Running OLS on the pooled *county-level* data with a variety of controls, Deschenes and Greenstone estimate that a 5 degree (F) increase in temperatures and an 8 percent increase in precipitation could reduce agricultural land values by \$75.1 billion ($t = 2.7$). When they include soil and socioeconomic covariates, however, the estimate changes to an increase of \$0.7 billion ($t = 0.0$), and when they include state fixed-effects as well the estimate changes

[11] They also include state-by-year fixed effects to control for aggregate yearly shocks by state. Their basic concern is that there could be long term time trends in weather at the region or state level; including the state-by-year effects will flexibly control for these trends.

to an increase of \$110.8 billion ($t = 4.7$).[12] OLS estimates therefore vary widely depending on the inclusion or exclusion of certain covariates.

Running models that include county and state-by-year fixed effects, Deschenes and Greenstone estimate that a 5 degree (F) increase in temperatures and an 8 percent increase in precipitation could increase agricultural profits by \$0.7 billion ($t = 1.7$).[13] Controlling for various covariates or experimenting with different specifications has minimal impact in this model – the estimated effect always remains close to zero and statistically insignificant. Note that the dependent variable has now shifted from land values to profits, since random year-to-year fluctuations in weather should have minimal effect on long term land values. Translating into land values using a 5 percent discount rate, the 95% confidence interval ranges from -\$2 billion to \$30 billion. This is much tighter than the range reported above for the cross-sectional models (which doesn't even account for sampling variation), and it rules out strong negative effects. In this sense, moving to the fixed effects specification seems to give more stable estimates that are a priori more credibly identified.

The downside to the fixed effects model, however, is that it only estimates the causal effect of *short-term* climate fluctuations on agricultural profits. There are good reasons to believe that the effects of a long-term increase in temperatures would be markedly different than the effects of a short-term increase. On the one hand, farmers might engage in adaptations in the long run that are unprofitable in the short run. This could reduce any negative effects that we observe in the short term. On the other hand, some short-run adaptations, e.g. drawing down reservoirs or storing crops, are not feasible in the long run. Furthermore, factors such as changes in the distribution of pests can take years to materialize. It is thus possible that the effects of long-term climate change are substantially worse than the effects of short-term climate fluctuations.[14] In many applications, including this one, it

[12] Note that even with state fixed effects, there is still some cross-sectional variation because the cross-sectional unit is the county.

[13] Deschenes and Greenstone's preferred estimate uses climate predictions from the Hadley 2 climate change model rather than the "benchmark" 5 degrees/8 percent scenario. Using the Hadley model, the estimated effect changes to \$1.3 billion.

[14] For additional discussion of the limitations, see Fisher, Hanemann, Roberts, and Schlenker (2012). In particular, there appears to be significant measurement error in Deschenes and Greenstone's coding of

can therefore be problematic to extrapolate too far off of estimates based on short-term, year-to-year variation.

4.6 Fixed Effects in Other Contexts

Although the focus of these notes is on panel data, the fixed estimator can be useful in other contexts as well. For example, Currie and Thomas (1995) use a fixed effects estimator to determine the effect of the Head Start program (an early intervention program available to children in poorer families) on test scores and health outcomes. Because children in poor families are more likely to enroll in the program, Currie and Thomas include mother fixed effects to control for any family-specific characteristics that affect all children similarly. The FE estimates thus use within-family variation – comparing outcomes for siblings that attended Head Start to outcomes for siblings that did not attend Head Start – to estimate the effects of Head Start.

Table 1 presents estimation results for both whites and African Americans. The first set of rows presents the coefficient from a bivariate regression of Peabody Picture Vocabulary Test (PPVT) scores on a Head Start indicator. Head Start appears to have a negative effect on whites (almost surely due to selection) and no effect on blacks. The second set of rows presents the coefficient from a regression of PPVT scores on a Head Start indicator, controlling for other covariates such as household income, mother's education, etc. The coefficient for whites changes to imply no effect of Head Start, while the coefficient for African Americans remains substantially unchanged. The third set of rows presents the coefficient from a regression of PPVT scores on a Head Start indicator, including mother fixed effects and some child-specific covariates (including household income at the time the child was three years old). Now the coefficient for whites is finally positive and significant, though for African Americans there is still no effect. Applying the fixed effects model has a substantial effect on the (white) estimates, even in comparison to an OLS regression with a rich set of covariates.

weather. This may cause attenuation bias in some of the specifications.

Table 1: Effects of Head Start on PPVT Scores

	Whites	Blacks
OLS Unadjusted	-5.62	1.04
	(1.57)	(1.22)
OLS Adjusted	-0.38	0.74
	(1.45)	(1.14)
Fixed Effects	5.88	0.25
	(1.52)	(1.36)

Source: Currie and Thomas (1995). Parentheses contain standard errors. Data are from NLSY.

5 Additional References

Fisher, T, M. Hanemann, M. Roberts, and W. Schlenker. (2012) "The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather: Comment." *American Economic Review*, 102:7, pp. 3749-60.