

Panel Data Data

Nathaniel Beck
Department of Politics
NYU
New York, NY 10012
nathaniel.beck@nyu.edu

Spring 2004

Panel Data (Especially Modeling “Effects”)

We not turn to panel data. This has large N (all asymptotics in N , fixed T , usually small (T may even be 2, usually though not always less than 5)).

Note this kills certain strategies which involve averaging over T time units, but opens up strategies which involve averaging over N .

Starting simple, we continue with the baseline model

$$y_{i,t} = \mathbf{x}_{i,t}\beta + \epsilon_{i,t}; \quad i = 1, \dots, N \quad (1)$$
$$t = 1, \dots, T$$

of yesterday. If this holds, then OLS is still fine. Unlike TSCS, we do NOT usually worry about spatial effects (is respondent 1007 related to respondent 1432?) and temporal effects are hard to model, since we have so few time points observed.

The threat that is typically dealt with is known as “unmodeled heterogeneity.” (Modeled heterogeneity are all the differences between $x_{i,t}$ and $x_{j,s}$ that lead to differences in $y_{i,t}$ and $y_{j,s}$.) Unmodeled heterogeneity is basically all the stuff that differentiates $\epsilon_{i,t}$ from $\epsilon_{j,s}$.

As is usual, focus is on heterogeneity in units; we could also focus on heterogeneity in time, or both. Since N is large, we can pick up heterogeneity in time pretty easily, as we shall see. Thus we usually worry about some unmodeled relationship between $\epsilon_{i,t}$ and $\epsilon_{i,s}$,

Obviously the best solution is to model the heterogeneity, that is, to find some x to explain it. If wishes were horses

Robust (clustered) standard errors

Since the observations for unit i are clearly not independent, OLS (and ML where we assume that the log of the likelihood is just the sum of the individual log likelihoods) is no longer optimal and its standard errors are no longer right.

OLS is still consistent. Maybe we can just fix the standard errors. This is done by a procedure first described by Huber, popularized (independently) by White, and well implemented in Stata.

The idea is easiest to see in the linear world (where we are today!), but is general to ML (remembering that OLS is ML). In ML, the regularity conditions imply that the OPG (outer product of the gradient) is equal to the Hessian.

In English: If you are walking down a mountain, first moving a few steps east, then a few steps south, is the same as moving the few steps south, then east, or of moving directly southeast the same amount (so long as your steps are REALLY SMALL).

For a simple cross-sectional ML estimator (with independent observations) its variance covariance matrix is given by

$$\widehat{\text{VCV}} = -\mathbf{H}^{-1} \sum_{i=1}^N \left[\left(\frac{\partial \ln L_i}{\partial \beta} \right)' \left(\frac{\partial \ln L_i}{\partial \beta} \right) \right] -\mathbf{H}^{-1} \quad (2)$$

where \mathbf{H} is the Hessian, the matrix of second partials.

Under the regularity conditions, the OPG is equal to the negative inverse of the Hessian, and so we get our usual formula that the variance of the ML estimator is the negative inverse of the Hessian. (This is the also the justification for BHHH, which estimates the VCV by the OPG.)

Now some like this “sandwich estimator” of the VCV in general, though, as Greene notes, if the regularity condition fails so as to make the sandwich necessary, who knows what is going on.

In the linear case, under heteroskedasticity, the Huber formula reduces to White’s famous “robust” estimator of the VCV,

$$\text{CovHC} = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_i (e_i \mathbf{x}_i)' (e_i \mathbf{x}_i) \right) (\mathbf{X}'\mathbf{X})^{-1} \quad (3)$$

where \mathbf{x}_i is the vector of exogenous variables for observation i . This ML formula simplifies to this on looking at the score and Hessian of linear ML.

This works for heteroskedasticity. What about panels. The similar Huber estimate of the “cluster robust” VCV is

$$\text{CovPH} = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^N \mathbf{u}_i' \mathbf{u}_i \right) (\mathbf{X}'\mathbf{X})^{-1}, \mathbf{u}_i = \sum_{t=1}^T e_{i,t} \mathbf{x}_{i,t} \quad (4)$$

Note: the Huber se's do not fix the problem of heterogeneity, but they do help fix the se's, AND THEY APPEAR TO DO NO HARM. Thus they should always be used here in preference to OLS se's, assuming you decide to use OLS. Trivial to do in Stata.

So far we have ignored heterogeneity from unit to unit. One way to model heterogeneity is to add effects to the model (unit specific intercepts).

Fixed effects

The simplest approach to modeling heterogeneity is to assume that each individual has his or her own specific intercept, f_i . Thus we have N separate parallel regression lines. The “fixed effects” model is

$$y_{i,t} = \mathbf{x}_{i,t} \beta + f_i + \epsilon_{i,t} \quad (5)$$

where all is like Equation 1 except for the varying intercepts.

This model is often called “analysis of covariance;” without the covariates, \mathbf{x} , this would be straight “analysis of variance.” (If we added a time dummy, d_t , we would have “two way analysis of variance” (or covariance with the \mathbf{x} terms included). It is also called the “within estimator.”

Note: Since T is small, adding a few g_t fixed time effects to the model is usually econometrically trivial and so we focus on the unit effects. As we have seen in the Meier data, ignoring temporal variation can have enormous consequences.

Note that N is large, so we will have a lot of effects to estimate. Also note that each effect is estimated with T observations, so we won't get a very good estimate of any single f_i . But we usually don't care about the effects *per se*, they are “nuisance parameters.”

LSDV

We could just estimate Equation 5 by OLS. In the old days (and even now), it was hard to estimate an OLS with thousands of parms, so we actually use a trick (this trick is implemented in Stata AREG). This trick is also theoretically critical (though of less practical importance!) since it allows us to avoid the “Neyman-Scott incidental parameters” problem.

Neyman and Scott showed (in 1948) that ML was inconsistent (or potentially inconsistent) if the model contained nuisance parameters such that the number of nuisance parameters grew asymptotically with then number of observations. Note that fixed effects model, the f_i are nuisance parameters that have this feature. Thus we have to worry that these do not cause inconsistent estimates of the β . (As noted above, consistent estimates of the f_i is close to an oxymoron.)

To get around the incidental parameter problem, we estimate the β condition on the fixed effects, which solves the Neyman-Scott problem. In general this is a hard thing to do, but for linear panel models, it is easy.

The trick is to “sweep out the unit effects” by taking deviations of all variables from the UNIT means. We thus form $y^*_{i,t} = y_{i,t} - \bar{y}_i$ where \bar{y}_i is the average y score in unit i . We similarly transform x and then regress y^* on x^* . The resulting regression has no unit effects.

LSDV - justification

While this makes intuitive sense or can just be seen as a linear transform of the model which leave the transformed estimates of β intact, it is also a trivial consequence of the standard Frisch-Waugh theorem on partitioned models. This theorem (a trivial consequence of the inverse of a partitioned matrix) tell us that we can estimate β in $y = x\beta + z\gamma$ by first regressing y and x (separately) on z and then regressing the residuals of y on all the separate residuals of the x regressions. (If you have forgotten this, Greene, ch. 6.4 has a very clear discussion.)

The notation here gets a bit dicey since we need to construct some $NT \times N$ matrices (which are all block diagonal). To avoid tedium, just write the LSDV model as

$$y = [\mathbf{D}, \mathbf{X}][\mathbf{f}, \beta]' + \epsilon \quad (6)$$

where the y and ϵ are stacked vectors (starting with the T observations for unit 1, etc.) and \mathbf{D} is a matrix of “dummy variables” \mathbf{f} is a vector of fixed effects.

The F-W theorem tells us we can estimate β by regressing y and each x on the set of dummies, \mathbf{d}_i , which produces the deviations of the $y_{i,t}$ and $x_{i,t}$ from their UNIT means. and then regress these deviated y on the deviated x . (Greene, p. 560-1 has the completely anal derivation, but is nothing more than what is here, once you remember your definitions of mean and residual making matrices).

Again, if you have a N such that your computer can estimate a model with N ind. vars, you could just do straight regression. (Since Stata limits you to regressions with 800 variables, you cannot estimate the Meier model with fixed effects without using AREG.)

Testing for fixed effects

Since we are doing OLS, the test for whether you need fixed effects, that is, a test of the null that all the units share the same intercept, against the alternative that they vary,

$$H_0 : f_i = f_j \forall i, j, H_1 : f_i \neq f_j \text{ for at least one } i \neq j \quad (7)$$

is the usual F test which compares the sum of squares in the LSDV regression and the regression with no dummies but a single constant.

(Practical note: while it does not matter, AREG estimates a model with no constant while a typical regression program wants a constant, so you have to drop f_1 (say) and then all then the constant term becomes f_1 and all other estimated effects are the difference between f_1 and the estimated effects. Since we usually don't care about the actual effects (if we did we would be in trouble, they are so poorly estimated) this doesn't matter, and any F test procedure will get this right, so long as you remember the numerator has $N - 1$ df.)

Should we include fixed effects?

If the F is significant, should we include the fixed effects. Depends on the gain and the cost. The gain is indicated by the change in the sum of squares. Note that with big $N \times T$, a small change in sum of squares may be statistically significant.

What are the costs.

1. LSDV kills all ind. vars. that don't vary over time (demographics). Cf. Yoon's paper for an example. Is this bad. Only if we care about the demographics. Note that the fixed effects are just like unmeasured demographics, where the demographics are so complicated that each person has own demographics.

2. LSDV is very inefficient, in that it estimates a LOT of parameters. It uses up $\frac{1}{T}$ of the degrees of freedom available. If $T = 2$ (or even 3), this is an enormous cost.

3. LSDV makes it hard for any slowly changing ind. vars. to survive, since they are highly co-linear with the effects. Thus imagine in PID is a covariate. It is not a demographic so may change, so not perfectly co-linear with the effect, but is unlikely to survive the inclusion of an effect.

4. So do not do LSDV without thinking about it, even if the F test is significant. The costs of LSDV are one reason that people use random effects.

5. We will deal with this later, but note that many of the arguments against LSDV hold for panels, not TSCS data. In particular, the T in TSCS data is large enough that you can get good estimates of the effects and item 2 is less relevant. 1 and 3 are relevant. For more

discussion of these issues, see Yoon (paper), Green, Kim and Yoon (IO) and Beck and Katz (IO), all on reserve.

Random effects

The alternative is called random effects. Instead of adding a dummy f_i , we take the “random effect”, α_i is a draw from some distribution (often normal), so the random effect models is just

$$y_{i,t} = \mathbf{x}_{i,t}\beta + \{\alpha_i + \epsilon_{i,t}\} \quad (8)$$

which is like Equation 1 except that $\text{VAR}(\epsilon_{i,t}) = \sigma_\epsilon^2$ and $\text{VAR}(\alpha_i) = \sigma_\alpha^2$ and we have to worry about the joint distribution of α and ϵ though we will solve all problems by assumption.

Obviously this has practical value, since we cut N extra params down to whatever is needed to characterize the new “error” process.

Note that if we have a big T (TSCS data), we will get a lot or repeated observations on whatever distribution generated α_i so fixed effects will be similar to random effects. For small T (panel), we need to assume a common distribution for all the α , since any unit’s draws will give us little information on the process if we assume separate parameters for each unit.

Advantages of Random Effects

The obvious disadvantage of random effects is that we have to make assumptions about the distribution. The obvious gain is the N df. But random effects also make some sense, particularly if we care about the effects. We saw the OLS estimates of the fixed effects are very bad. Consider a panel, and assume no covariates (doesn’t matter, if you have covariates, just take out those effects). So we have

$$y_{i,t} = f_i + \epsilon_{i,t} \quad (9)$$

What is the best estimate of f_i . Obviously \bar{y}_i , so $\bar{\epsilon}_i = 0$.

Consider a short panel of two waves and suppose there are no effects in reality (but we estimate Equation 9 anyway). Suppose y is approval of Clinton and suppose $\epsilon \sim N(0, 100)$. Given that, by chance a fair number of people will have two draws exceeding +10, and hence an estimated effect of at least 10, and a similar number on the negative side, even though by assumption there are no effects. The fixed effects estimator is unable to realize that these “effects” are just do to chance. But we would not want to overstate this: two draws of +100 would lead us to believe that this person likes Clinton more than the average person does.

Estimating the random effects, and in particular $\widehat{\alpha}_i$ is called a BLUP (best linear unbiased predictor) and is much better estimator of the effect than the corresponding \widehat{f}_i .

The BLUP “shrinks” \hat{f}_i back to the overall mean effect (in this case, zero). We will see lots more shrinkage estimators tomorrow. Since we as political scientists typically do not care about estimating α_i , I will not go through the calculation, which is a subset of tomorrow’s calculation (which I also won’t really go through). Note also that since we are estimating effects, we can (and should) produce standard errors of these estimates. Stata produces the estimates of the random effects, but not the standard errors; mixed estimation schemes, which we will discuss tomorrow, easily produce both. If you really care about estimating effects, use SPLUS.

Estimating Random Effects

For estimation we must ASSUME that the α and ϵ are independent. Such an assumption is commonplace. Without it we could not separate the two random processes.

We also assume, benignly, that both sets of errors are spherical.

We MUST also ASSUME, less benignly, that the x and α are independent (for the same reason that x and ϵ must be independent. This means that folks who are unusually rich should not be also more likely to like Clinton for unmeasured reasons. This assumption has generated a lot of literature.

In terms of estimation, note that the random effects model is just Equation 1 with a more complicated Ω . This complication comes from the error process now being $\alpha_i + \epsilon_{i,t}$ so the errors for a given unit are correlated with each other (each containing a common α_i).

Thus, the variance of the total “error” for observation (i, t) is just $\sigma_\alpha^2 + \sigma_\epsilon^2$, the covariance of the errors for two different observations from the same unit is σ_α^2 and the errors from different units are independent.

The relative weight of the random effects α and the individual observation errors (ϵ) is given by the estimated variances of those two processes. If $\sigma_\alpha^2 = 0$, there are no random effects.

There is a standard Lagrange multiplier test for whether the random effects are necessary (that is, $H_0 : \sigma_\alpha^2 = 0$). This is just based on estimating the model under the null (Equation 1 and then seeing if the individual residuals are related to the average unit residuals. This test is easy to do in Stata (xttest0).

We can estimate Equation 8 by maximum likelihood, where we break the likelihood down into N independent likelihoods, one per unit. This ML approach is quite complicated, and based on a lot of tricks to break the T -variate normal into something more tractable. The discussion of this estimator is always based on some cite to some much earlier work, and some suggestion of some great econometrician. It may or may not be good. It is not commonly used. It is implemented in Stata (without much documentation, so one might

worry). It usually doesn't vary much from the FGLS estimator which everyone uses. If you want to pursue, there is some discussion and references in both Hsiao and Baltagi.

FGLS estimation of RE models

To be more formal, the VCV matrix of the “augmented” errors ($\alpha + \epsilon$) is

$$\mathbf{V} = \begin{pmatrix} \Omega & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Omega & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Omega & \dots & \mathbf{0} \\ & & & \ddots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \Omega \end{pmatrix} = \Omega \otimes \mathbf{I}_N \quad (10)$$

where

$$\Omega = \begin{pmatrix} \sigma_\epsilon^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\epsilon^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ & & & \ddots & \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\epsilon^2 + \sigma_\alpha^2 \end{pmatrix} \quad (11)$$

Given this, it is straightforward to do GLS, since we just need the inverse “square root” of this matrix, which has a nice simple form since the inverse square root of an identity \otimes a matrix is just an identity \otimes the inverse square root of that matrix, and the Ω matrix has a nice inverse square root.

As is typical, GLS then consists of transforming all observations, with

$$y_{i,t}^* = y_{i,t} - \theta \bar{y}_i \quad (12)$$

where

$$\theta = 1 - \frac{\sigma_\epsilon}{\sqrt{T\sigma_\alpha^2 + \sigma_\epsilon^2}} \quad (13)$$

(and similarly, of course, for the ind vars).

Note that this is just like LSDV except instead of subtracting the unit means from each observation, we subtract some fraction of those unit means ($0 \leq \theta \leq 1$).

Note if $\theta = 1$ we get LSDV. Except for a silly case, this only happens at $T \rightarrow \infty$, that is, as are panels become TSCS data (only issue is T , not N .)

If $\theta = 0$ then pooled GIS is pooled OLS (Equation 1), which happens if $\sigma_\alpha^2 = 0$, that is, no effects (all drawn from a distribution with a mean and variance of zero).

This shows why fixed and random effects differ. In the panel world, if you have random effects, then the fixed effects model overstates the size of the fixed effects (by attributing some random variation that should be in the ϵ to the effect, as we have seen). The GLS weighting for random effects gets this right (assuming random effects are right!).

Another way to see this is that GLS is a weighted combination of OLS (a regression of the original data) and LSDV (the regression subtracting off group means), with θ determining the weight put on each of the two components.

FGLS estimation of random effects

Of course we know neither of the two variances in the formula for θ , so need to use FGLS with estimated variances.

The trick of FGLS estimation is that there are a number of consistent estimators of β in Equation 8. Thus, OLS is consistent, since Equation 8 is just an ordinary linear model with a complicated (heteroskedastic and correlated) error process. LSDV would also be consistent. A third estimator is the so-called “between estimator.” This just averages all the observations for unit i and then regresses \bar{y}_i on \bar{x}_i . None of these three are fully efficient or produce correct standard errors, but all produce consistent estimates of β .

Note that LSDV, by taking deviations from unit means, also sweeps out the random effect. (If $y_{i,t} = \alpha_i + \text{stuff}$, then $y_{i,t} - \bar{y}_i = \text{stuff} - \text{mean unit stuff}$, the α_i is taken out.) Thus the residuals from the LSDV equation can be used to estimate σ_ϵ^2 (with careful correction for df if you like unbiasedness).

The between estimator, use N observations on

$$\bar{y}_i = \bar{x}_i + \epsilon_i^B \tag{14}$$

where the error term, ϵ^B contains both ϵ and α . Thus $\epsilon_i^B = \bar{\epsilon}_i + \bar{\alpha}_i = \bar{\epsilon}_i + \alpha_i$ since every “error” contains an α_i term. Since the two error processes are independent (by assumption), we have

$$\text{Var}(\epsilon^B) = \sigma_\epsilon^2 + \sigma_\alpha^2 \tag{15}$$

which can be combined with the LSDV estimate of σ_ϵ^2 to yield an estimate of σ_α^2 . These two estimates are then put back into the GLS, formula, and we have FGLS (with a slightly wrong standard error, since we estimate the variances rather than know them for sure).

In practice you may get a negative estimated σ_α^2 , which is annoying, but in such cases we think that σ_α^2 is small, at best, and so OLS is fine (there are no effects).

Fixed or random effects

The random effects estimator differs from the fixed effects estimator insofar as $\frac{\sigma_{\epsilon}^2}{\sigma_{\epsilon}^2 + T\sigma_{\alpha}^2}$ differs from zero. As T gets large this term, however, goes to zero. Thus for TSCS data the difference is usually small.

Fixed effects are right if we want to make inferences conditional on the observed units. That is, if we think of sampling a new set of OECD political economy data, will Germany remain the same. If so, a fixed effect for Germany is correct

In a large N panel study, we don't care about the units, and we would think of sampling new units. Thus prediction is going to be unconditional on the sampled units. Here random effects are appropriate.

If we have a random effects model, we can make forecasts conditional on the sample. This will be the fixed effects, shrunk back towards zero, with the degree of shrinkage determined by the homogeneity of the data. Why should we shrink towards zero. Think of estimating a fixed effect with $T=3$. Sometimes will be positive on average just by chance. The shrinkage corrects for that. With a large T , the possibility of getting a large number of positive error draws by chance declines, and so the estimated fixed effect is good. REMEMBER TO THINK ABOUT PANEL VS TSCS DATA!

So fixed effects are fine for TSCS data. Note that fixed effects may not be necessary or may not be a good thing (Przworski and Teune say a good model should contain no proper names, and the fixed effects are exactly proper names). Also we have a problem with fixed effects if some of the $x_{i,t}$ do not vary by time. But in TSCS models they should! (You do not want to explain change in unit i over time by something that is invariant over time in that unit.

Note that if a model has a unit variable that does not change over time, we cannot estimate a fixed effects model, since the temporally stable variable is just a multiple of the unit fixed effects. We can, however, still estimate a random effects model. This is not a silly way to proceed.

Hausman test for fixed vs random effects (sort of)

The big assumption of the random effects model is that the effects are uncorrelated with the errors. We know that the assumption that errors are uncorrelated with the ind vars. is the most critical regression assumption; since the random effects are part of the error for GLS, the assumption is equally critical here. And more or less equally untestable. If you think that effects are correlated with the ind vars., you can't use random effects.

(This also tells us why the random effects estimates of β are usually pretty similar to the OLS estimates. Random effects just changes the error process. Unless σ_{α}^2 is large in comparison to σ_{ϵ}^2 , FGLS and OLS will not differ by that much, though estimated standard errors will differ (there is some difference in the estimated β for efficiency reasons. But the efficiency gains are usually not that great, at least for small ϵ_{α}^2 .)

One can do a Hausman test to see if random effects are “okay in some sense.” I prefer that formulation to the usual one, that is, that the effects are uncorrelated with the ind vars, because there are so many reasons that the Hausman test can lead to rejection of the null, and I find it hard to figure out what to make of this rejection. (Failure to reject the null means that random effects are fine, but they are similar to the fixed effects, though they are more efficient. It is rejection of the null that is less conclusive.)

A Hausman test compares two estimators. Under the null, both are consistent, but one is more efficient; under the alternative, the more efficient of the two becomes inconsistent but the less efficient remains consistent. Thus if the null is okay, the two estimators should be similar; divergence indicates rejection of the null.

This gives the test statistic

$$(\hat{\beta}^{GLS} - \hat{\beta}^{LSDV}) \hat{\Sigma}^{-1} (\hat{\beta}^{GLS} - \hat{\beta}^{LSDV}) \quad (16)$$

where $\hat{\Sigma}$ is the difference of the estimated covariance matrices from the two estimators.

A large value of the statistic indicates rejection of the null, that is, the typical interpretation at least, that the effects are correlated with the ind vars. and we should not use GLS (but rather LSDV). I only infer that the estimators differ.

For completeness, the statistic has χ^2 distribution with K df, where K is the number of elements of β . The Hausman test for panel data is implemented in `xthaus` in Stata.

There are many reasons that LSDV will yield different estimates of β , especially for small T and covariates that do not move very much. In fact, I would expect, *a priori*, that the Hausman test will reject the null for many political science models. What we do with that rejection is unclear to me.

Costs and Benefits

What do we get from panels that we don't get from the analysis of repeated surveys (the gain in temporal dimension over simple cross-sections is obvious).

- Get a handle on unmodeled heterogeneity (if effects work)
- Get better information on previous behavior/attitude

What are the costs

- COSTS
- Panel Mortality
- Could be serious, though Bartels says not in NES
- Panel fatigue
- Priming

Issue - how will this work with internet surveys?

Panel Data - Fixed Effects - Hurwicz Bias

If we have a dynamic panel, with

$$y_{i,t} = \mathbf{x}_{i,t}\beta + \rho y_{t-1} + \alpha_i + \epsilon_{i,t} \quad (17)$$

With fixed effects we underestimate ρ . This is known as Hurwicz or Nickell bias. It is the direct analogue that a simple time series autoregression underestimates the autoregressive coefficient (Hurwicz). This bias is of order $\frac{1}{T}$, and so in normal time series studies is not a problem.

But in panel studies it is of the same order, that is, for short panels, very serious.

Where does this bias come from. In a simple regression, we “sweep out” the constant term by subtracting \bar{y} and \bar{x} from the data, and then regress the centered data. Note that in a simple autoregression x is just y_{t-1} . When we subtract \bar{y} , we are subtracting a term which is the average of T terms, and one of those terms is y_{t-1} . Thus there is a slight correlation between the independent variable (the lagged y) and the transformed error process. This correlation gets smaller as T gets larger.

The same problem arises with panel data with fixed effects. Now we are sweeping out not \bar{y} but \bar{y}_i . If this average is only over two or three $y_{i,t}$, then the correlation between $y_{i,t-1}$ and the transformed error process is not trivial. Thus we have the bias fully documented by Nickell.

Solution - First Differencing and IV

If we difference Equation 17 we get

$$\Delta y_{i,t} = \rho \Delta y_{t-1} + \Delta \mathbf{x}_{i,t}\beta + \Delta \epsilon_{i,t} \quad (18)$$

The problem is that $\Delta y_{i,t-1}$ contains $y_{i,t-1}$ which is by construction correlated with $\epsilon_{i,t-1}$ and hence with $\Delta \epsilon_{i,t}$.

Thus OLS is inconsistent.

But we can use instruments for $\Delta y_{i,t-1}$. Both $y_{i,t-2}$ and $\Delta y_{i,t-2}$ are correlated with $\Delta y_{i,t-1}$ and uncorrelated with any ϵ (assuming the ϵ are temporally independent).

Thus we can use IV estimation (assuming $T > 2$!!!!).

Wawro (manuscript) has details and various improved estimators, some of which are implemented in various Stata xt commands.

Panel Data - Random Effects

The problem with random effects is the compound "error term" is $\alpha_i + \epsilon_{i,t}$. Since α_i is correlated with $y_{i,t-1}$ by construction, we have an endogeneity problem.

Unlike single time series, the results critically depend on assumptions about the initial $y_{i,0}$. (In single time series, this matters, but disappears as T grows. When $T = 2$ or 3 , the initial condition matters!)

One can start by first differencing and then using the fixed effects IV estimator.

One can then use these estimated parameters to estimate the variance of the random effects.