

POLS 509: The Linear Model, *Lecture # 12*

Eric Reinhardt

Apr 14, 2005

Techniques for Panel Data

Apr 14 (Th) & Apr 19 (Tu): (12) Techniques for panel data.

- Wooldridge, 426-483.
- Greene, 283-338.
- Homework # 10 due Tu Apr 19.
- Homework # 11 distributed.

1 Outline

- 1) Concepts & Common Problems
- 2) To pool or not to pool?
- 3) Error components models (fixed effects, random effects, Hausman test)
- 4) FGLS vs. OLS with panel-corrected standard errors

2 Concepts

Consider a dataset consisting of a set of cross-sectional entities, e.g., countries, with repeated observations on each cross-section over time, e.g., each year. For example,

Country	Year	i	t	$X_{i,t}$	$Y_{i,t}$
France	1980	1	1	25	15
France	1981	1	2	29	22
France	\vdots	\vdots	\vdots	\vdots	\vdots
France	2005	1	26	45	3
Canada	1980	2	1	60	12
Canada	1981	2	2	62	10
Canada	\vdots	\vdots	\vdots	\vdots	\vdots
Canada	2005	2	26	79	2
Brazil	1980	3	1	12	27
Brazil	1981	3	2	11	38
Brazil	\vdots	\vdots	\vdots	\vdots	\vdots
Brazil	2005	3	26	25	17

This is a **panel** dataset, with three cross-sections (denoted by i), each observed for 26 separate time periods (t).

There tends to be significant heterogeneity across countries or panels, in the intercepts or error variance, among other possibilities. Also, because each cross-sectional case is observed sequentially, the fact that unobserved sources of variation tend to remain from period to period introduces the possibility of serially correlated errors.

Imagine a very different kind of dataset, which merges a number of (different) cross-sections drawn independently over time, so that each entity does **not** get observed repeatedly. The Pooled National Election Studies 1972–2004 would be an example, since each year of the survey, a different set of respondents is selected for interviewing.

Respondent	Year	X_i	Y_i
Schwimmer, D.	1980	.	.
Aniston, J.	1980	.	.
\vdots	\vdots	\vdots	\vdots
LeBlanc, M.	1980	.	.
Alexander, J.	1990	.	.
Seinfeld, J.	1990	.	.
\vdots	\vdots	\vdots	\vdots
Louise-Dreyfus, J.	1990	.	.

This is quite different from the “panel” dataset above. Wooldridge calls it an **independently-pooled cross-section** dataset. While the variables are not likely to be identically distributed from period to period, there is little prospect of serial correlation since no one respondent or cross-section is observed repeatedly. Such datasets present fewer problems as a result. Hence today let’s stick with the **panel dataset** context.

Let’s make one last distinction:

Call the number of cross-section entities in our panel, e.g., countries, N . The country is identified as $i \in \{1, 2, \dots, N\}$.

The number of time periods per cross-section is T , with any particular period $t \in \{1, 2, \dots, T\}$.

So the total number of observations in the sample is $N \times T$.

The distinction to make is between panel datasets with far more cross-section units than time periods per unit ($N > T$)¹ and, conversely, those with far more time periods per unit than cross-sections ($T > N$).² As we will see, life is generally easier with the latter kind of dataset.

3 Should You Pool Your Data?

Consider our basic regression model, subscripted for the panel data context:

$$Y_{i,t} = \beta_0 + \beta_1 X_{1,i,t} + \beta_2 X_{2,i,t} + \dots + u_{i,t}$$

As stated, this model assumes that the values of the Y -intercept/constant and all of the coefficient β s are **the same for all cross-sectional units**. This can be a very strong assumption.

¹Example: country-years in a panel of developing countries for 1985-1999, with $N \approx 140$ and $T = 15$.

²Example: monthly data on economic performance for 12 OECD countries for the period 1970-2005 ($N = 12$, $T = 432$).

Consider an example, from the file `r:\lectures\CWS.dta`, which contains the Comparative Welfare State Dataset compiled by Huber, Ragin, and Stephens (and updated by others). This file is structured as a panel dataset, with 17 developed countries observed for 15 years (from 1960 through 1994). Estimate the regression

$$\text{cpi}_{i,t} = \beta_0 + \beta_1 \text{neocorp}_{i,t} + \beta_2 \ln \text{pc}_{i,t} + \beta_3 \text{capital}_{i,t} + \beta_4 \text{leftcab}_{i,t} + u_{i,t}.$$

That is, a country's inflation is a function of its level of "neocorporatism" (coordination of wage bargaining, and state-market relations, at the peak level by business and labor confederations), its per capita income, its openness to international capital flows, and the partisan orientation of its government.

What we are implicitly assuming here is that neocorporatism, partisan orientation of the government, etc., affect inflation **the same way for every country (and time period) in the analysis.**

Is this a reasonable assumption? Well, we can test its validity by generalizing the equation with interaction terms to allow for differences across unit categories (or before and after a common break in the time series for all) in the values of the β s.

Let's say we suspect that the explanatory variables may have different consequences (i.e., coefficients) in the continental European countries (for which $\text{CE}=1$) than in the other countries in our sample.

To perform a **Chow Test** of structural stability of the coefficients across two subsamples, we estimate a generalized regression

$$Y_{i,t} = \beta_0 + \beta_1 X_{1i,t} + \beta_2 X_{2i,t} + \theta_0 G_{i,t} + \theta_1 G_{i,t} X_{1i,t} + \theta_2 G_{i,t} X_{2i,t} + u_{i,t},$$

where $G_{i,t}$ is a dummy (0 or 1) denoting which subsample group observation i, t lies in, and the θ s are the coefficients of the resulting interaction terms.

The test statistic is then just the F statistic on the null hypothesis that the θ s are collectively equal to zero. We can allow the two subsamples' intercepts to differ by limiting the F test to the null that $\theta_1, \theta_2 = 0$. And, while this test is not robust (i.e., could be wrong) to heteroskedasticity and/or autocorrelation, we can estimate the generalized regression with appropriately robust SEs, and then test the same nulls using a Wald test (with the same exact Stata command, `test`) with the appropriately corrected variance-covariance matrix

NOTE: The validity of pooling is often not questioned in practice. Authors tend to seek a generalizing theory and they don't want to allow for anything less general by including such interaction terms. Theory is the best guide, but we can see in our example that the assumption that a political process works the same across all cases can be heroically demanding of the data.

4 Error Components Models

A special, and commonly-used, approach to allowing some heterogeneity across cross-sectional units in a pooled data analysis, is to permit the constants or Y -intercepts to vary across each unit, even if the coefficients are fixed. These are called **error components models**, since they are based on a small modification of our typical OLS equation:

$$Y_{i,t} = \mathbf{X}_{i,t}\boldsymbol{\beta} + a_i + u_{i,t}.$$

The a_i parameter has potentially different values across the cross-sections, but is the same over time within a given unit (because it doesn't subscript by t as well). The idea of error components is that we can imagine a composite error term

$$e_{i,t} = a_i + u_{i,t}$$

which, if put into our equation above, would yield $Y_{i,t} = \mathbf{X}_{i,t}\boldsymbol{\beta} + e_{i,t}$, a simple regression equation.

So the composite error can be broken into components, one being an additive intercept term unique to each cross-section, the other being a true error following our OLS assumptions.

There are two alternative ways of viewing the a_i , however.

4.1 Fixed Effects OLS

The simplest approach is to treat the a_i as constants. We simply include a set of dummy variables, one for each cross-sectional unit, leaving one out as the reference category (whose constant is simply β_0 as a result). The issue then is that a portion of the composite error term $e_{i,t} = a_i + u_{i,t}$ then does not vary over time within a panel unit, yielding serial error correlation. But we can rationalize this fixed effects approach since

$$Y_{i,t} = \mathbf{X}_{i,t}\boldsymbol{\beta} + a_i + u_{i,t}$$

hence

$$\bar{Y}_i = \bar{\mathbf{X}}_i\boldsymbol{\beta} + a_i + \bar{u}_i,$$

a regression with just one observation for each cross-section, using the averages of each variable within each time series. The latter is known as the **between-effects estimator** (which, if it alone is your primary approach, demands large N in your panel).³ Subtract each side of this between-effects equation from each side of the first, to strip out the a_i , and we get

$$\begin{aligned} Y_{i,t} - \bar{Y}_i &= \mathbf{X}_{i,t}\boldsymbol{\beta} - \bar{\mathbf{X}}_i\boldsymbol{\beta} + a_i - a_i + u_{i,t} - \bar{u}_i \\ &= \boldsymbol{\beta} (\mathbf{X}_{i,t} - \bar{\mathbf{X}}_i) + u_{i,t} - \bar{u}_i. \end{aligned}$$

Since $E(u_{i,t}) = 0$ by assumption, this leaves us with just OLS, which is BLUE and estimates the same $\boldsymbol{\beta}$. This latter equation is called a **time-demeaned transformation** and it yields the **within estimator**, another

³In Stata, obtain the between-effects estimator with `xtreg y x..., i(i) be`.

name for **fixed effects**. This focuses on the internal variation across time *within* each cross-sectional unit, unlike the between-effects approach, which focuses solely on the variation across cross-sectional units.

Practically, we can estimate fixed effects models crudely by creating a set of dummy variables, one for each panel unit i , and including them on the RHS of an OLS regression. By adding a literal variable for every a_i , what's left of the composite error term is just the true error $u_{i,t}$.

In Stata, you can `tab i, gen(i_)` and then `reg y x... i_*`. Since this form includes a constant, Stata will arbitrarily drop one of the dummies you included, which serves as the reference category, without loss of generality. You can use the `, robust` option or the `newey2...`, `lag(m)` command layered over this procedure, to obtain heteroskedasticity-consistent or heteroskedasticity- and m -order autocorrelation-consistent standard errors. Test the hypothesis that all the fixed effects are collectively equal to zero, i.e.,

$$H_0 : \mathbf{a} = \mathbf{0}$$

(where the a_i s are represented in an $i \times 1$ column vector), with `testparm i_*`.

You can somewhat more elegantly run `areg y x..., abs(i)` (which can take the `robust` option but not `newey2` at present) or `xtreg y x..., i(i) fe` (which does not allow a robust SE option but which has other diagnostic commands specially allowed for it).

Final note: you can easily (albeit manually) add time-fixed effects in addition to the panel-fixed effects, a la

$$Y_{i,t} = \mathbf{X}_{i,t}\boldsymbol{\beta} + a_i + b_t + u_{i,t}.$$

If you are working with what Wooldridge calls independently-pooled cross-section data (as in a pooled set of surveys drawn separately over time), while there is little reason to suspect that $a_i \neq 0$ for any cross-section

(i.e., respondent) i , you probably *do* have different distributions of the variables across *time*. In this case, the conventional technique is to include time-specific fixed effects, without panel-specific fixed effects. (Although if your survey has many respondents in each of a variety of different regions/countries/conceptual categories, you may want to include dummies for each of those regions/countries/categories, if not for each respondent individually. Such dummies are conceptually the same as “fixed effects.”)

Given that it is a version of OLS, when its assumptions are satisfied, fixed effects panel estimation is BLUE, as even a small-sample property.

Problems:

1. Parameters that don't vary across time within a unit will be dropped and absorbed into that unit's fixed effect.
2. Uses up many degrees of freedom, esp. with a large- N (# of cross-sections) model, so less efficient than desired in some cases. But, it is still unbiased even if a_i is correlated with the X s, even if this increases the SEs due to multicollinearity. It is accordingly desirable to have large T for a fixed N !

4.2 Random Effects GLS

Let's alternatively adopt the much more demanding assumption that the a_i are truly uncorrelated with \mathbf{X} (over all time periods). Then our composite error term, and not just its true error component $u_{i,t}$, meets our traditional OLS assumption. In particular, $Cov(\mathbf{X}_{i,t}, a_i) = 0$ for all i, t .

The idea of random effects is that the panel-specific error component (or intercept) a_i is itself a random variable rather than being a (fixed) constant, composed of a common mean (say, α) and a random element ε_i specific to each cross-section. If our generic error components model is

$$Y_{i,t} = \mathbf{X}_{i,t}\boldsymbol{\beta} + a_i + u_{i,t},$$

we decompose

$$a_i = \alpha + \varepsilon_i.$$

a_i 's distribution is the same for every cross-section i , so that $E(a_i) = \alpha$. because $E(\varepsilon_i) = 0$. Say that

$$\begin{aligned} \sigma_a^2 &= \text{Var}(a_i) \text{ (the same for all cross-sections!)} \\ \text{and } \sigma_u^2 &= \text{Var}(u_{i,t}) \text{ (also constant across all observations).} \end{aligned}$$

We further assume that the covariance between any two cross-sections' ε_i , and the covariance between ε_i and $u_{i,t}$, are all zero, conditional on \mathbf{X} .

But this causes a problem of serial error correlation in the composite error term

$$e_{i,t} = a_i + u_{i,t} = \alpha + \varepsilon_i + u_{i,t},$$

since the α are the same over time.

The idea is simple: let's strip that out using GLS, just as with any type of autocorrelation. Given the variances of the components as noted above,

$$\text{Corr}(e_{i,t}, e_{i,t'}) = \frac{\sigma_a^2}{(\sigma_a^2 + \sigma_u^2)},$$

which is the amount of serial error correlation between time points t and t' . We can compute weights based on this correlation:

$$\lambda = 1 - \sqrt{\frac{\sigma_u^2}{(\sigma_u^2 + T\sigma_a^2)}}$$

And we transform our observations of Y and $\mathbf{X}_{i,t}$ by subtracting out by those weights in a time-demeaned expression like so:

$$Y_{i,t} - \lambda \bar{Y}_i = \beta_0 (1 - \lambda) + \beta_1 (X_{1,i,t} - \lambda \bar{X}_{1,i,t}) + \dots + (a_i + u_{i,t} - \lambda \alpha).$$

Presto! A GLS estimator!

Like all GLS estimators, this requires true knowledge of the panel-specific error variance σ_a^2 and the common error variance σ_u^2 . We don't know these in practice, so we rely on estimators of them which are **consistent** if not unbiased as $N \rightarrow \infty$ with a fixed T . (Remember, this is the opposite data condition than the one that fixed effects is most efficient in, but here the property at issue is consistency rather than efficiency.) These estimators are based on squared fitted OLS residuals, just as we have seen in prior approaches. This then produces a feasible GLS, or FGLS, estimator for random effects.

Want a more intuitive explanation of the random effects GLS equation above? It is actually just a λ -weighted average of the between-effects and fixed-effects estimators. Take a look.

Issues to remember:

1. RE is more efficient than FE if its assumptions hold, particularly if N is large and T is very small.
2. But it yields **biased coefficient estimates** if the key assumption, that the random panel-specific intercepts are not correlated with the independent variables \mathbf{X} , does not hold.
3. Like other cases of efficiency-vs-consistency tradeoffs, we can test whether the RE assumption is appropriate in any given context with the Hausman test.

4. Yet the RE and FE estimators converge as $\lambda \rightarrow 1$, which is true as $T \rightarrow \infty$ (see the formula for λ above) and, to a lesser extent, as σ_a^2 increases relative to σ_u^2 . Conversely, the potential finite-sample bias in the RE estimator is more severe to the extent that $\lambda \rightarrow 0$.

To estimate a random effects model in Stata, we simply run

```
xtreg y x..., i(i) re
```

4.2.1 Breusch-Pagan LM Test for Random Effects

We may want to test the collective validity of including random effects in the model, just as earlier we tested the collective significance of the fixed effects with an auxiliary F test. The way to do this is to use a (different) Breusch-Pagan test with the null

$$H_0 : \sigma_a^2 = 0.$$

If the null is true, then the random effects equal zero and need not be included. (Plain OLS would then be preferred.)

The test statistic is

$$LM = \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^N (T\bar{e}_i)^2}{\sum_{i=1}^N \sum_{t=1}^T e_{i,t}^2} - 1 \right]^2,$$

which follows a $\chi^2(1)$ distribution. (Its derivation relies on maximum likelihood, so I won't go into it.)

In Stata, after running the RE model, use `xttest0`. For a version robust to serial error correlation, download and use `xttest1`.

4.3 Hausman Test

The approach is the same as discussed under the autocorrelation topic. The idea: test the null that the RE (i.e., FGLS) coefficients are consistent. If they are, then they will be the same as the known-to-be-consistent FE (i.e., OLS) coefficient estimates. If they are very different from each other, weighted appropriately by their variance-covariance matrices, then we reject the null.

[your way of running FE here]

```
estimates store consistent
xtreg y x..., i(i) re
estimates store efficient
hausman consistent efficient
```

4.4 Autocorrelation in the FE and RE Context

Going beyond the scope of this course, let me just say that there are a variety of additional models based on the error components structure and least squares estimation that also allow for autocorrelation. Stata incorporates the approach of Baltagi and Wu in its `xtregar` command....

5 Another GLS Model and OLS with PCSEs

5.1 Formal Definition of Some Concepts

Let's start by defining some particular concepts alluded to earlier, working from a basic regression model generalized and subscripted for our panel dataset context:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{a} + \mathbf{U} \\ \text{or } Y_{i,t} &= \mathbf{X}_{i,t}\boldsymbol{\beta} + a_i + u_{i,t}, \end{aligned}$$

where a_i varies across units but not within a unit observed across time.

Different Intercepts We need $a_i = a_j$, for all $i, j \in \{1, 2, \dots, N\}$. If not, our cross-sectional units have different Y -intercepts, or regression constants. [This is the issue addressed in both the FE and RE approaches, so let's assume away these cross-sectional effects for the rest of this discussion.]

Panel Heteroskedasticity We have panel (or, as Greene calls it, "group-wise") heteroskedasticity if $E(u_{i,t}^2) \neq E(u_{j,t}^2)$ but $E(u_{i,t}^2) = E(u_{j,t'}^2)$, so that

$$E(u_{i,t}^2) = \sigma_i^2,$$

a common error variance over time within a cross section but potentially different error variances *across* units.

Contemporaneously Correlated Errors This applies if we have $E(u_{i,t}u_{j,t}) = E(u_{i,t'}u_{j,t'}) \neq 0$ but $E(u_{i,t}u_{j,t'}) = 0$, so that

$$E(u_{i,t}u_{j,t}) = \sigma_{i,j}$$

with other covariances equalling zero. This means that the errors in any two units, observed at the same period, are correlated, but that there is no correlation of errors across those two units when observed at *different* points in time. [spatial autocorrelation]

Unit-Specific Autocorrelation This obtains if

$$u_{i,t} = \rho_i u_{i,t-1} + v_{i,t},$$

where $v_{i,t}$ are bound by the usual CLRM assumptions about errors, but the composite error term $u_{i,t}$ exhibits a unique degree of (here, just first-order) serial correlation for each cross-sectional unit.

Common Autocorrelation A simpler case of the above obtains if

$$u_{i,t} = \rho u_{i,t-1} + v_{i,t}.$$

The difference is that ρ is not subscripted here.

5.2 XTGLS

A version of FGLS has been developed to focus on these latter problems (Greene 320-33). Let's assume first-order unit-specific rather than merely common autocorrelation. Now, remember, the GLS estimator for a known error variance-covariance matrix $\mathbf{\Omega}$ is just

$$\hat{\beta}_{GLS} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1} (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{Y}).$$

But we don't really know $\mathbf{\Omega}$ and thus we must rely on a consistent estimator of it.

In practice, the strategy is to run auxiliary OLS regressions in sequence, first, a la Prais-Winsten, to estimate the ρ_i , then once again on the transformed (now, FGLS) variables, to estimate the contemporaneous error correlation, then a third time on the appropriately-transformed variables, yielding the final FGLS estimates.

In Stata, use `xtgls y x..., i(i) t(t) panels(corr) corr(psar1)`. You can relax the contemporaneous correlation part of this and keep just the panel heteroskedasticity by changing `panels(corr)` to `panels(het)`. You can relax the panel-specificity of the autocorrelation and turn that into a common autocorrelation parameter by changing `corr(psar1)` to `corr(ar1)`. If the panel is unbalanced (has missing data), you may need to use the `force` option (which tells Stata to treat two sequential nonmissing observations as t and $t + 1$, even if they are not really adjacent in your dataset).

One **key limitation** here is that T must be at least as big as N . There are $N(N + 1)/2$ contemporaneous covariances, each of which you must estimate with only NT observations. Beck and Katz (1995) show by simulation that you really need $T > 3N$, since in practice these FGLS SEs become biased downwards by a factor of 4, 5, 6, or more as you approach $T = N$. Moreover, this FGLS approach also has the well-known problem of inconsistency (i.e., biased coefficient, and not just SE, estimates) in small samples.

5.3 PCSEs

Given these limitations, Beck and Katz argued for using OLS with appropriately corrected SEs. Their **panel-corrected standard errors** (PCSEs) are based on the consistent estimator

$$\hat{\sigma}_{i,j} = \frac{\sum_{t=1}^T \hat{e}_{i,t} \hat{e}_{j,t}}{T},$$

derived from a first OLS regression and its fitted residuals $\hat{e}_{i,t}$.

The elements $\hat{\sigma}_{i,j}$ compose an $N \times N$ matrix of contemporaneous error covariances across the cross-sectional units. If we have a large T , this estimate of the matrix, Σ , becomes more accurate, you can see.

The actual PCSEs are then derived from

$$VarCov(\beta) = \left(\sum_{i=1}^N \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^N \sum_{j=1}^N \hat{\sigma}_{i,j} \mathbf{X}'_i \mathbf{X}_j \right) \left(\sum_{i=1}^N \mathbf{X}'_i \mathbf{X}_i \right)^{-1}.$$

OLS with PCSEs yields unbiased and consistent coefficient estimates and consistent SEs. They perform well in simulations even when the heteroskedasticity and error correlation structures are not present, i.e., OLS assumptions about errors apply. It is, however, a bit less efficient than the FGLS approach above, but that disadvantage is likely outweighed by the downward bias in the FGLS SEs which occurs at low T/N ratios.

You can add autocorrelation to this by running the Prais-Winsten FGLS transformation on the OLS estimates, and then reporting PCSEs on the resulting weighted OLS results.

In Stata, `xtpcse y x...` gives the basic estimates (without autocorrelation correction). Add the option `, corr(psar1)` to run the Prais-Winsten transformation in addition. Try it and see that the resulting SEs are indeed larger than those of a corresponding XTGLS result.

NOTE: What place would the Hausman test have in deciding which approach to use? Well, the Beck-Katz critique, and the PCSEs they propose, bear on the **standard errors**. The problems we have seen in the past with FGLS have not been about SEs, but about bias in the coefficients. That potential problem, of course, still applies here, regardless of whether we use PCSEs or regular OLS SEs (or other robust SEs). So we can address that issue with the Hausman test. But the Hausman test does *not* speak to the appropriateness of OLS/PCSEs vis-a-vis FGLS in terms of any bias in the FGLS SEs. That is an issue you must address solely by virtue of the guidelines and limits in FGLS noted above. Beck and Katz, as the developers of PCSEs, highlight their simulation evidence that PCSEs almost always dominate the FGLS approach, which at best is only marginally more efficient while potentially massively biased in terms of its SEs.