

# Longitudinal (Panel and Time Series Cross-Section) Data - Day 4 - Random Coefficient Models

Nathaniel Beck  
Department of Political Science  
University of California, San Diego  
La Jolla, CA 92093  
beck@ucsd.edu  
<http://weber.ucsd.edu/~nbeck>

Summer, 2001

## Modeling heterogeneity

While will do both TSCS and panel, different issues relevant for different data.

Start with TSCS until further notice

Some of this works for panel, but must check!

In PS, assumption of perfect heterogeneity (pooling, eg. Equation 1 on Day 1) is usually not remarked on.

Or one can assume full heterogeneity (by unit, this is standard time series) and no one would remark on it.

Can we do something in between - units are alike, and so their parameters are related, but they are NOT identical (save for the covariates).

How assess heterogeneity?

How model?

## Classic Test

The classic test for pooling is to take Equation 1 of Monday as the null with the alternative being complete heterogeneity, that is,

$$H_0 : y_{i,t} = \mathbf{x}_{i,t}\beta + \epsilon_{i,t} \quad (1)$$

$$H_1 : y_{i,t} = \mathbf{x}_{i,t}\beta_i + \epsilon_{i,t} \quad (2)$$

that is,  $H_0 : \beta_i = \beta$ .

The test of this is the standard F test, that is, take the difference of SSE's of the two models, correct for df, and put in the usual F ratio.

Note that if, say,  $k = 5$  and  $N = 20$ , there are 100 df in the numerator, which is a lot.

(Note: the F test may suggest homogeneity of all coefs because most coefs are homogeneous and they few that aren't can't overcome the loss of degrees of freedom from those that are.)

If you have one variable (call is  $z$  of interest, with the others ( $\mathbf{x}$ ) being more or less controls, you might consider only testing the heterogeneity of that parameter, and then if it is heterogeneous, estimating the model

$$H_0 : y_{i,t} = \mathbf{x}_{i,t}\beta + z_{i,t}\gamma_i + \epsilon_{i,t} \quad (3)$$

Obviously this need not be limited to one  $z$ . (Lagged dependent variables and controls for the worldwide economy might be good candidates for homogeneity *in some models*.)

(Note: You may reject the basic null because one or two coefficients vary enough and the  $F$ -test is not very conservative)

One also might worry that the  $F$  with so many degrees of freedom in the numerator will not pick a very parsimonious specification (the critical  $F$  value approaches one as the df in both numerator and denominator get large, that is, the  $F$  test becomes maximize  $\bar{R}^2$ ).

While this would take us far afield, a criteria that picks more parsimonious models and that is liked by applied worked is the BIC or Schwarz criteria. (All criteria are  $f(\text{SSE}) + \text{penalty for lack of parsimony}$ , penalty is roughly  $\frac{k}{N}$  for  $F$ -like criteria, much higher (roughly  $\frac{k \ln N}{N}$ ) for Bayesian criteria.

## Cross-Validation

One also might want to ask the data which countries do not fit the pooled specification. One way to do with would be via cross-validation.

In simplest form, c-v estimates a model leaving out one obs at a time, and then compares the prediction for that obs with the actual obs. This is the cross-sectional analogue of out-of-sample time series forecasting.

For large data sets, the cross-validated errors converge on the residuals (since dropping one obs hardly changes the estimated parms). For small data sets simple c-v is quite useful.

But one can also do “leave-out-k(%)” as well as “leave-out-one” c-v. This works nicely for TSCS data, where we can just leave one country out at a time, predict it, and then examine the “forecast” errors. (CV is also used for model selection, and we could use this leave out one unit CV for this purpose also.)

This is done in the table for a model of Garrett’s. Typical mean absolute forecast errors range from 1.2 to 2 (the unit is percent growth in GDP), except for Japan, which has a forecast error of 3.2% of GDP. Thus clearly Japan fits the basic specification much less well than any other OECD nation.

Table 1: Out of sample forecast errors (OLS) by country for Garrett model of economic growth in 14 OECD nations<sup>a</sup>, 1966–1990

Country	Mean absolute error
US	1.9
Canada	1.7
UK	1.7
Netherlands	1.6
Belgium	1.6
France	1.2
Germany	1.4
Austria	1.3
Italy	1.7
Finland	2.0
Sweden	1.2
Norway	1.5
Denmark	1.7
Japan	3.2

<sup>a</sup>No unit effects

## Random Coefficients Models

The random coefficients model (RCM) is an interesting compromise between assuming complete homogeneity and complete heterogeneity. This model is the same as the Bayesian hierarchical model (Western, 1998).

The RCM is a compromise between estimating the fully pooled and a fully unpooled estimate (Equations 1 and 2. The latter is a separate OLS estimate for each unit.

There is not enough data for the latter (that is separate OLS estimations will have huge standard errors), but the former requires the very strong assumption of complete pooling. The RCM model uses the idea of “borrowing strength” (Rubin, Gelman et al., etc.)

## Shrinkage

Shrinkage estimators are due to Charles Stein (1950's)

He showed that if your criteria was mean squared error loss (that is, the expected sum of squared errors in your parameter estimates), that a shrinkage estimator was better than OLS (though the shrinkage estimator is biased for any single parameter, if one cares about that).

The estimator is a linear combination of the OLS estimator and zero, with the combination parameter chosen cleverly. (It is intimately related to ridge regression, which also shrinks, but a little less cleverly).

Best explanation is Judge and Bock. They show that the Stein estimator is just a cleverer “pretest” estimator.

A pretest estimator tests (usually)  $H_0 : \beta = 0$  and if do not reject, estimates  $\hat{\beta} = 0$ , if reject, take OLS estimate. This is common practice and not totally crazy, if true  $\beta$  is large, you are very likely to reject the null.

But it does have an odd discontinuity.

Judge and Bock show that Stein is just a pretest estimator (that is, a linear combination of 0 and OLS  $\hat{\beta}$ , with the weight being a function of the  $F$  statistic on the test of the null. Thus if you clearly reject, you end up almost entirely with the OLS estimate, if you reject with a really small  $F$ , you end up almost entirely with  $H_0$ .

One reason people don't like Stein (among some silly ones) is why should you shrink to zero, what is special about zero. (Actually you could shrink to anything and get better MSE performance, the  $F$  keeps you from shrinking very much to something very stupid

## Borrowing Strength

The Bayesians (following Rubin) have pushed shrinkage, this time to a group mean, under the rubric “borrowing” strength. Suppose you have a bunch of crumbly estimates of some score (say improvement in SAT scores at various test centers, with each test center having a small number of test takers).

You could assume all scores are from same distribution, and take overall mean. This has low variability, but makes very strong assumption.

It take separate scores, no assumption of pooling but no statistical power.

Suppose you had five centers with scores 1,1,-1,-1 and 10. You might believe the 10 center is better than the others, but you might also think that 10 was the result of some lucky draws (remember the discussion of random effects).

Efron and other “empirical Bayesians” suggest shrinking each score back to the overall mean, with the degree of shrinkage being proportional to the  $F$  statistic on the null that all 5 centers have the same underlying distribution (mean and variance).

Rubin is Bayesian, so he suggests using a “prior” that the centers are all the same, with the data used to update that. As a modern Bayesian, he uses a “gentle prior.”

Theil and Goldberger showed how you could incorporate non-sample information into classical regression, the so-called “mixed model” (though it is hard to make theoretical sense of how a classicist can have non-sample information).

In any event, one either ends up with “empirical Bayes” (where the prior is given by the empirical variation among the means) or modern Bayes (which is Bayes with a gentle or relatively uninformative prior). How much different these are in practice is unclear.

## Back to TSCS

For TSCS the obvious thing to do is to estimate each unit by OLS, and then shrink the estimates back to the overall mean, with the degree of shrinkage being given by the  $F$  statistic we got from testing Equation 1 vs 2. In Beck and Katz (2001) we show that for TSCS data you simply do not shrink very much, that this form of shrinkage is very close to unit by unit OLS unless you have a very small  $T$  or a lot of random coefficients (which we think you should not!). Note that for *panels* that  $T$  is small!

## The formal RCM model

Formally the model we are considering is:

$$\begin{aligned}y_{i,t} &= \mathbf{x}_{i,t}\beta_i + \epsilon_{i,t} \\ \beta_i &\sim N(\beta, \mathbf{\Gamma}) \\ E[\beta_i - \beta | \mathbf{x}_{i,t}] &= 0 \\ E[\epsilon_{i,t} | \mathbf{x}_{i,t}] &= 0 \\ E(\epsilon_{i,t}\epsilon_{j,t}) &= \begin{cases} \sigma_i^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} .\end{aligned}\tag{4}$$

When we consider ML estimates we will further assume:

$$\epsilon_{i,t} \stackrel{\text{ind}}{\sim} N(0, \sigma_i^2).$$

Note that we could allow the errors to have a more complicated error structure, but this would cloud the central issues.

### Some useful things

$$\nu_i = \beta_i - \beta.$$

Clearly  $\nu_i \sim N(0, \Gamma)$ . We can re-write the model for the  $y_{i,t}$ 's by substituting as:

$$\begin{aligned} y_{i,t} &= \mathbf{x}_{i,t}\beta + (\epsilon_{i,t} + \mathbf{x}_{i,t}\nu_i) \\ y_{i,t} &= \mathbf{x}_{i,t}\beta + w_{i,t}. \end{aligned} \tag{5}$$

with  $w_{i,t}$  as our new composite error term. The first part of  $w_{i,t}$  is the standard stochastic part of a regression model. The second term is the error associated with how far a particular unit's  $\beta_i$  is from the overall mean  $\beta$ .

It will also be convenient to stack the observations by unit instead of considering individual observations. Define

$$\mathbf{y}_i = \begin{bmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,T} \end{bmatrix} \quad \mathbf{x}_i = \begin{bmatrix} \mathbf{x}_{i,1} \\ \mathbf{x}_{i,2} \\ \vdots \\ \mathbf{x}_{i,T} \end{bmatrix} \quad \boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \\ \vdots \\ \epsilon_{i,T} \end{bmatrix} \quad \mathbf{w}_i = \begin{bmatrix} w_{i,1} \\ w_{i,2} \\ \vdots \\ w_{i,T} \end{bmatrix}.$$

### OLS

If it were the case that  $\mathbf{w}_i$  were suitably well behaved, and, in particular, that  $E[\mathbf{w}_i | \mathbf{x}_i] = 0$ , then we could estimate Eq 5 by OLS, which would give us a consistent, although possibly not efficient estimate of  $\beta$ . In fact, we have already made enough

assumptions to ensure this. To see this note:

$$\begin{aligned}
 E[\mathbf{w}_i | \mathbf{x}_i] &= E[\boldsymbol{\epsilon}_i + \mathbf{x}_i v_i | \mathbf{x}_i] \\
 &= E[\boldsymbol{\epsilon}_i | \mathbf{x}_i] + E[\mathbf{x}_i v_i | \mathbf{x}_i] \\
 &= 0 + \mathbf{x}_i E[v_i | \mathbf{x}_i] \\
 &= 0.
 \end{aligned}$$

The last step is true because we have assumed in the basic setup that  $\beta_i$  are mean independent of the  $\mathbf{x}_{i,t}$ . Therefore, there cannot be a systematic relationship between the unit's average  $\mathbf{x}_{i,t}$  and  $\beta_i$ . This assumption would fail if units with high values of some regressors also had larger  $\beta_i$ .

Consider, for example, a comparative model of annual government spending as a function of revenue among other things. It might be the case that governments that are better at raising funds are also more likely to have higher spending — i.e., have a large  $\beta_i$ .

### Pooled OLS as RCM estimator

We thus have our first candidate RCM model: *pooled OLS*.

We should note, however, our OLS estimates would not be efficient for the overall mean because it is not using the full structure of our model. This can be seen by examining the covariance matrix of the pooled OLS estimate defined by Eq 5:

$$\begin{aligned}
 E[\mathbf{w}_i \mathbf{w}_i'] &= E[(\boldsymbol{\epsilon}_i + \mathbf{x}_i v_i)(\boldsymbol{\epsilon}_i + \mathbf{x}_i v_i)'] \\
 &= \sigma_i^2 I_T + \mathbf{x}_i \boldsymbol{\Gamma} \mathbf{x}_i' \\
 &= \boldsymbol{\Pi}_i.
 \end{aligned}$$

So for the full sample (stacking observations) the covariance matrix of OLS estimate of is

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Pi}_1 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Pi}_2 & \mathbf{0} & \dots & \mathbf{0} \\ & & \vdots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Pi}_N \end{bmatrix} \tag{6}$$

$$= I \otimes \boldsymbol{\Pi}_i \tag{7}$$

Where  $\otimes$  is the Kronecker product. If we knew  $\boldsymbol{\Gamma}$ , as well as  $\sigma_i^2$ , we would know  $\boldsymbol{\Pi}_i$  and therefore  $\boldsymbol{\Omega}$ . Estimation would be relatively straightforward generalized least squares (GLS). We will explore this more below.

## Unit by unit

At the other extreme from complete pooling would be to run OLS on each unit. We will refer to this estimate as *unit by unit OLS* and denote the estimate from the  $i$ 'th unit by  $\mathbf{b}_i$ . Clearly, this estimate borrows the least strength from other units, and will be consistent even if there is some correlation structure between  $\beta_i$  and  $\mathbf{x}_i$ . Asymptotically (as  $T \rightarrow \infty$ ) this would be the preferred estimator because we can recover  $\beta_i$  regardless of the relationship between units (which is only indirectly of interest in practice). However, in finite samples, this estimator may have very large sampling variance.

## Stein-Rule

As was mentioned before, they were designed to be smooth versions of pre-test estimators (which have very complicated sampling distributions) based on the F test for parameter homogeneity. Accordingly, the weighting between the pooled and unpooled estimators is a function of this test statistic. Formally the Stein-rule estimator for unit level parameters is

$$\hat{\beta}_i = \frac{c}{F}\hat{\beta} + \left(1 - \frac{c}{F}\right)\mathbf{b}_i$$

where  $\hat{\beta}$  and  $\mathbf{b}_i$  are defined above,  $F$  is the statistic for testing the null hypothesis of equality of the  $\beta_i$  and  $c$  is a constant. Judge and Bock (1978:190–195) suggest that the optimal value for this constant is:

$$c = \frac{(N - 1)k - 2}{NT - Nk + 2}, \quad (8)$$

where  $k$  is the number of regressors.

## Generalized Least Squares

An alternative shrinkage estimators can be considered in the generalized least squares framework. Recall that the estimate of the overall mean from the pooled OLS was inefficient because it does not use all of the information in the structure of the model. A GLS estimate of  $\beta$  would be:

$$[\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{X}\boldsymbol{\Omega}^{-1}\mathbf{y}$$

Where  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]'$  and  $\mathbf{y} = [y_1, y_2, \dots, y_N]'$ . In general, we will not know  $\boldsymbol{\Gamma}$  but will have to estimate it. We will turn to how to estimate it shortly, but for now assume we have some consistent estimate of  $\boldsymbol{\Gamma}$ , say  $\hat{\boldsymbol{\Gamma}}$ . We can then construct the feasible generalized least square estimate (FGLS) as follows:

$$\tilde{\beta} = [\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}]^{-1}\mathbf{X}\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y} \quad (9)$$

where

$$\begin{aligned}\hat{\Omega} &= I_N \otimes \hat{\Pi}_i \\ &= I_N \otimes (\hat{\sigma}_i^2 I_T + \mathbf{x}_i \hat{\Gamma} \mathbf{x}_i').\end{aligned}$$

There is an alternative formulation of  $\tilde{\beta}$  as a weighted function of the unit by unit OLS estimates ( $\mathbf{b}_i$ ) which gives us a bit more insight into the GLS estimators. Clearly  $\mathbf{b}_i$  is consistent for  $\beta_i$  (and  $\beta$ ) by the same assumptions and argument we used to show that OLS using the observations from all units was consistent for the estimate of the overall mean  $\beta$ . Given standard results we can show:

$$\begin{aligned}\text{Var}(\mathbf{b}_i) &= (\mathbf{x}_i' \mathbf{x}_i)^{-1} \mathbf{x}_i' \mathbf{\Pi}_i (\mathbf{x}_i (\mathbf{x}_i' \mathbf{x}_i)^{-1}) \\ &= V_i + \mathbf{\Gamma},\end{aligned}$$

where

$$V_i = \sigma_i^2 (\mathbf{x}_i' \mathbf{x}_i)^{-1}.$$

Then the GLS estimator given in ( 9) can be written as:

$$\tilde{\beta} = \sum_{i=1}^N \mathbf{W}_i \mathbf{b}_i \tag{10}$$

where

$$\mathbf{W}_i = \left\{ \sum_{i=1}^N [\mathbf{\Gamma} + V_i]^{-1} \right\}^{-1} [\mathbf{\Gamma} + V_i]^{-1}$$

The GLS estimate, then, is a weighted average of the unit by unit OLS estimates. The weights are such that units that have smaller variance of their estimates, perhaps because they have a larger  $T$  or just fit better, are given more weight.

It will also be useful to have an explicit formula for the variance of  $\tilde{\beta}$ . Given Eq 10 and the fact that the  $\mathbf{b}_i$  are independent, the variance is straight forward to derive.

$$\begin{aligned}\text{Var}(\tilde{\beta}) &= \sum_{i=1}^N \mathbf{W}_i \text{Var}(\mathbf{b}_i) \mathbf{W}_i' \\ &= \sum_{i=1}^N \mathbf{W}_i [V_i + \mathbf{\Gamma}] \mathbf{W}_i'.\end{aligned} \tag{11}$$

We can now turn to the estimations of the parameters of interest,  $\beta_i$ . The best linear unbiased predictor of  $\beta_i$  that comes from the GLS framework is:

$$\begin{aligned}\hat{\beta}_i &= [\mathbf{\Gamma}^{-1} + \hat{\mathbf{V}}^{-1}]^{-1} [\mathbf{\Gamma}^{-1} \tilde{\beta} + \hat{\mathbf{V}}^{-1} \mathbf{b}_i] \\ &= A_i \tilde{\beta} + [I_k - A_i] \mathbf{b}_i\end{aligned}\tag{12}$$

where  $A_i = [\mathbf{\Gamma}^{-1} + \hat{\mathbf{V}}_i^{-1}]^{-1} \mathbf{\Gamma}$ . That is, the estimate is a weighted average between the pooled and unit by units. The weights are given by the relative precision of our two estimates. We can justify  $\hat{\beta}_i$  is two ways. From the classical perspective, it minimizes mean squared error given the setup in Eq 4. It is in that sense it is “best”. However, it can also be justified for the Bayesian perspective. If we were interested in estimating the  $\beta_i$  and we had a prior of the form,  $\beta_i \sim N(\beta, \mathbf{\Gamma})$ , this would be our posterior mode estimate. Further, if we are going to estimate  $\beta$  and  $\mathbf{\Gamma}$  from observed data (as we will below), we could call this an empirical Bayesian estimate. Depending on how we estimate  $\mathbf{\Gamma}$ , we will get different RCM estimators. We now turn to these.

It will be useful to define  $\text{Var}(\hat{\beta}_i)$  in order to do statistical testing. In general, since  $\hat{\beta}_i$  is estimated by a linear combination we can use standard results to show:

$$\begin{aligned}\text{Var}(\hat{\beta}_i) &= A_i \text{Var}(\tilde{\beta}) A_i' + [I_k - A_i] \text{Var}(\mathbf{b}_i) [I_k - A_i]' + \\ &\quad [I_k - A_i] \text{Cov}(\tilde{\beta}, \mathbf{b}_i) A_i' + A_i \text{Cov}(\tilde{\beta}, \mathbf{b}_i) [I_k - A_i]'\end{aligned}\tag{13}$$

This can be simplified by noting that  $\text{Cov}(\tilde{\beta}, \mathbf{b}_i) = \mathbf{W}_i \text{Var}(\mathbf{b}_i)$ . This is true because the  $\mathbf{b}_i$ 's are independent; the only term in the sum that makes up  $\tilde{\beta}$  that has positive covariance with a particular  $\mathbf{b}_i$  is  $\mathbf{W}_i \mathbf{b}_i$ .

## Problems with standard estimators

The obstacle for GLS is that we do not know either  $\mathbf{\Gamma}$  or  $V_i$ , hence we can not use Eq 10 to estimate the overall mean parameter or Eq 12 to estimate  $\beta_i$ . However, as is often true with GLS models we can use a two-step procedure.

We first consider an estimator originally suggested by Swamy (1971) to estimate  $\mathbf{\Gamma}$  and  $V_i$ . In the first step we use some consistent, but inefficient, estimator to estimate  $\beta$  and the  $\beta_i$ 's. We then use these preliminary estimates to estimate the variance parameters. In this case, we run OLS unit by unit to estimate  $\mathbf{b}_i$ . We then estimate  $\hat{V}_i$  by its usual estimate:

$$\begin{aligned}\hat{V}_i &= s_i^2 (\mathbf{x}_i' \mathbf{x}_i)^{-1} \\ s_i^2 &= \frac{\mathbf{e}_i' \mathbf{e}_i}{T - k},\end{aligned}$$

where  $e_i$  are standard OLS residuals and  $k$  is the number of regressors.

The question is how to estimate  $\Gamma$ ? If we could directly observe  $\beta_1, \dots, \beta_N$ , we could use the  $N$  draws to construct an estimate of the covariance matrix in the usual fashion:

$$\tilde{\Gamma} = \frac{1}{N-1} \left( \sum_{i=1}^N \beta_i \beta_i' - N \overline{\beta \beta'} \right),$$

where  $\bar{\beta}_i$  is the mean of the  $N$  observed  $\beta_i$ . We note that any such estimate of  $\Gamma$  will improve as  $N$  get large. Formally  $\tilde{\Gamma}$  will converge in probability to  $\Gamma$  using standard assumptions.

While this is important to panel analysts, who can allow for asymptotics in  $N$ , it is of cold comfort to TSCS analysts who must work with fixed, and typically not very large,  $N$ s. This issue is muted a bit if we take an empirical Bayesian perspective.

The point of the empirical Bayes approach is not to estimate  $\Gamma$ , per se, but to use it in forming of a prior to improve our estimates of  $\beta_i$ . Clearly there is information in the data about  $\Gamma$  even when  $N$  is fixed.

Again the problem is we do not observe  $\beta_i$ ; we have, instead, only a noisy estimates of the  $\mathbf{b}_i$ . So while we might consider just substituting  $\mathbf{b}_i$  for  $\beta_i$  in the definition of  $\tilde{\Gamma}$ , this would lead us to over estimate the amount of variation in  $\beta_i$  since much of the variation in the  $\mathbf{b}_i$ s is caused not by “real” parameter variability but purely by sampling error.

In finite samples we would expect the  $\mathbf{b}_i$ 's to differ. We can correct for this sampling variability by noting that the  $\text{Var } \mathbf{b}_i = V_i + \hat{\beta}$ . Swamy suggested that a plausible estimator of  $\Gamma$  is

$$\hat{\Gamma} = \frac{1}{N-1} \left( \sum_{i=1}^N \mathbf{b}_i \mathbf{b}_i' - N \overline{\beta \beta'} \right) - \frac{1}{N} \sum_{i=1}^N \hat{V}_i, \quad (14)$$

where  $\bar{\beta}$  is the mean of the  $\mathbf{b}_i$ 's. Thus the *Swamy estimator* plugs in the estimate for  $\Gamma$  into the formula for  $\hat{\beta}_i$ .

There is, however, a problem with this estimator: in finite samples  $\hat{\Gamma}$  (defined by Eq 14) need not be positive definite, a necessary requirement for it to be a well defined covariance matrix.  $\hat{\Gamma}$  may not be positive definite because we are subtracting off the mean  $\hat{V}_i$ , which can be large in finite samples. Recall that  $\hat{V}_i$  is the estimated sampling variability of the OLS estimate for unit  $i$ . If, for example,  $T$  is small, as it usually is in practice, we would expect  $\hat{V}_i$  to be large and to dwarf the effect of the parameter heterogeneity estimated by the first term.

The question is how to insure that our estimate of  $\Gamma$  is positive definite? Hsiao's suggestion, building on Swamy, is to drop the second term in the estimate of  $\Gamma$ , which seems to be the accepted practice. Thus we get the *Hsiao estimator* by plugging this estimate of  $\Gamma$  into Eq 12. The rationale for this is asymptotic. The first term of Eq 14 is of  $O(1)$  whereas the

second term is  $O(\frac{1}{NT})$ . In words the first term does not vanish as either  $N$  or  $T$  gets large since it is the estimate of the “true” parameter variability. The second term is sampling variability, so as  $T$  get large our estimate  $\mathbf{b}_i$  converge to their true values  $\beta_i$ , so the second term in Eq 14 vanishes *asymptotically in T*.

Note that this fix is not correct in finite samples as it will tend to overestimate  $\Gamma$ . The interesting question is how badly does this affect the estimate of  $\Gamma$  and does this cause any problems in the estimates of  $\beta_i$ ? Since these are problems in finite samples we will have to assess the claims using Monte Carlo simulations.

As alluded to above, we might be concerned that Hsiao estimate of  $\Gamma$  will be too large. If we look at the formula for  $\hat{\beta}_i$ , the larger is  $\hat{\Gamma}$ , the less we shrink. We might want to err the other way. So another possible estimator which we will refer to as *BKK* (for Beck-Katz kludge) is

$$\hat{\Gamma} = \max \left[ 0, \frac{1}{N-1} \left( \sum_{i=1}^N \mathbf{b}_i \mathbf{b}_i' - N \overline{\beta \beta'} \right) - \frac{1}{N} \sum_{i=1}^N \hat{V}_i \right]$$

That is, we use the Swamy estimator that corrects for sampling when it is positive definite; if not we set  $\hat{\Gamma}$  to zero (which means we are using the fully pooled OLS estimate of  $\beta_i$ ). This will have a complicated sampling distribution, but may work well in practice since it errs towards greater pooling.

## Bayesian and Maximum Likelihood Estimation

An alternative approach to two step FGLS estimate is a direct maximization of the likelihood. This can form the basis of either classical (i.e., maximum likelihood) or Bayesian analysis. The log likelihood defined by our RCM model can be written as:

$$\begin{aligned} \mathcal{L}(\beta_i, \sigma_i, \beta, \Gamma) = & K - \frac{T}{2} \sum_{i=1}^N \ln(\sigma_i^2) - \frac{1}{2} \sum_{i=1}^N \frac{1}{\sigma_i^2} (\mathbf{y}_i - \mathbf{x}_i \beta_i)' (\mathbf{y}_i - \mathbf{x}_i \beta_i) \\ & - \frac{N}{2} \ln |\Gamma| - \frac{1}{2} \sum_{i=1}^N (\beta_i - \beta)' \Gamma^{-1} (\beta_i - \beta) \quad (15) \end{aligned}$$

where  $K$  is some constant. In practice, however, this will be very difficult to directly maximize.

From a Bayesian prospective, the likelihood is combined with priors to generate posterior distributions of the parameters. In this RCM model (also called a hierarchical model in the

Bayesian literature), the priors are specified only for  $\beta$ ,  $\Gamma$ , and  $\sigma_i^2$ ; these then imply priors for the unit level parameters. In order to calculate the full posterior distribution in this model in general one would have to use numerical methods.

If we only wish to estimate the mode of this distribution (which is all ML does), then Smith (JRSS-B, 1973) gives a set of equations that define the mode of the posterior when the prior on  $\Gamma^{-1}$  is a conjugate Wishart with independent conjugate inverse  $\chi^2$  distributions as the priors for the  $\sigma_i^2$ . The resulting equations are:

$$\tilde{\beta}_i = \left( \frac{1}{\sigma_i^2} \mathbf{x}_i' \mathbf{x}_i = \Gamma^{-1} \right)^{-1} \left( \frac{1}{\sigma_i^2} \mathbf{x}_i' \mathbf{x}_i \mathbf{b}_i + \Gamma^{-1} \tilde{\beta} \right) \quad (16)$$

$$\tilde{\beta} = \frac{1}{N} \sum_{i=1}^N \tilde{\beta}_i \quad (17)$$

$$\hat{\sigma}^2 = \frac{1}{T + v_i + 2} \left[ v_i \lambda_i + (\mathbf{y}_i - \mathbf{x}_i \beta_i)' (\mathbf{y}_i - \mathbf{x}_i \beta_i) \right] \quad (18)$$

$$\tilde{\Gamma} = \frac{1}{N - k - 2 + \delta} \left[ R + \sum_{i=1}^N (\beta_i - \beta)' \Gamma^{-1} (\beta_i - \beta) \right] \quad (19)$$

where  $k$  is the number of regressors and  $R$ ,  $v_i$ , and  $\lambda_i$  are parameters that correspond to the prior. To get uninformative priors, we can set  $v_i = 0$ ,  $\lambda_i = 0$ , and choose  $R$  to be a diagonal matrix with small values, say 0.001. These equations can be solved iteratively with initial starting values from the unit by unit runs. The iterations continue until the parameter values converge.

## Monte Carlo results

Monte Carlo results (reported in Beck and Katz, 2001) show that the Hsiao estimator has bad properties. Maximum likelihood (and our kludge) both do okay. Surprisingly, pooling works well, even where an  $F$ -test would reject the null of pooling.

Consequence: do NOT use RCM routines in Limdep or Stata (xtrrhh). Splus maximum likelihood routines (Bates and Pinhero) probably okay, though hard to see exactly what they do. Empirical Bayes okay.

But perhaps should impose a stronger prior of pooling than given to us by data; with large  $T$ , do not pool as much as comparativist might want.

## Simplifications

Others (eg Bates and Pinhero) have SPLUS code that estimates Equation 4 by either ML or REML (the RE conditioning on OLS estimates of the error parameters).

Western and others use modern Bayes with a gentle prior.

But note that no one actually estimates the full Equation 4; it is just too complicated. Note that the error process is a  $k$ -variate normal, where in our work  $k$  is often 5 or more. This is an impossible estimation task, and in all my experience either ML takes forever, or failing that, returns garbage.

There are simplifications that are sensible, and still much better than the complete simplification of full pooling.

The first is to assume the random coefficients are independent normals, so no big covariance matrix to estimate. Thus we have  $\beta^k \sim N(0, \sigma_k^2)$  and  $\text{Cov}(\beta^k, \beta^{k'}) = 0, k \neq k'$ . This cuts down the estimation problem enormously.  $k$  here refers to the  $k$ 'th ind var (whether super or sub script).

Note that Western makes this assumption without even really mentioning it. It is our only change to actually estimate an RCM model in practice.

The other idea, like with the first  $F$  test, is to fix as many of the nuisance variables as possible. Not all variables really need to be random. The estimation problem is hard: simplify.

## Improved estimation

At that point we can estimate the model either using Western's modern Bayesian methods or full ML.

Both methods are computationally intensive.

The full ML model is done by Pinhero and Bates and implemented in Splus. See their BOOK for details.

The Western approach is MCMC (Winbugs). See Jackman's lectures later in the course for this.

## Modeling the random coefficients

The RCM can be made more useful by allowing the  $\beta_i$  to be functions of other unit variables,  $z_i$ , which allows for modeling differential effects as a function of differing institutions.

(Note: the  $z$  are time invariant, so they only measure properties of units.)

This is particularly important in comparative politics, where we might expect that the effect of some  $x$  on the dependent variable is contingent on structural features that vary from country to country. As an example, the Garrett model asserts that the effect of having a left government is contingent on the type of labor bargaining in each country.

We can then write:

$$\beta_i = \mathbf{z}_i\gamma + \beta + \alpha_i. \quad (20)$$

(with the  $\beta$  term usually just picked up as a constant in the  $z$  vector).

Substituting Equation 20 into Equation 4, we see that this model is just an interactive model with random coefficients on the linear terms only. Formally:

$$y_{i,t} = \mathbf{x}_{i,t}(\mathbf{z}_i\gamma)\beta + \{\mathbf{x}_{i,t}\alpha_i + \epsilon_{i,t}\} \quad (21)$$

(which contains all the linear  $x$ 's assuming Equation 20 contains a constant term, as it will).

Thus you estimate this model by simply rewriting to enter the multiplicative terms and estimate like any other RCM. The linear  $x$  enter through a constant term in the  $z$  vector.

Note: this is exactly the hierarchical or multilevel model. The multilevel model is just an RCM with interactive terms!

## Summary

- Pooled OLS not to bad
- Do we care about estimates of  $\beta_i$  or  $\beta$ ?
- Other ways to understand diversity (region)
- Test -  $F$ -test and variants
- Cross-Validation
- Practical - DO NOT USE STANDARD HSAIO ESTIMATOR AS IN LIMDEP OR STATA
- ML as in Splus probably okay
- Only reason to use MCMC is if easier or faster
- Should we be "real" Bayesians?

## Panel

The argument for RCM is much more powerful for panels. Here it is used all the time, is just called the multilevel model.

With large  $N$  it is hard to assess heterogeneity, and not at all obvious that we care if respondent 1312 is different from respondent 1103. So usually just go ahead and use multilevel model.

Lots of ways (and software) for doing multilevel.

See Jones and Steenbergen AJPS Workshop piece (2000) for a discussion.

Very useful in policy analysis, say schools, where one has data on individuals, schools, districts, etc. (Also, don;t forget to keep on modeling time, as in the Meier data.)