Advanced MLE

Introduction to Longitudinal Data
July 24, 2007

Introduction

As the course description suggests, this is a course on the analysis of longitudinal data. Despite the apparent obviousness of this, please note a few things:

- It is *not* a course on time-series analysis, econometric or otherwise.
- The substantive examples will come primarily – but not exclusively – from *political science*, though I'll try to be a bit catholic whenever possible.
- It is an *advanced* course; please drop the class and take something else if you're afraid of math.

All that said, I think the course materials will be very useful to most of you. The emphasis is on data that have both *cross-sectional* and *temporal* variation; most commonly, this takes the form of repeated measurements over time on multiple units of observation. We'll talk more about these two dimensions of variation in a bit...

Some (Non-Technical) Terminology

Throughout the course, I'll try to be consistent in terminology and notation. So:

- We'll use the words "unit," "units," and/or "units of observation" to refer to the individual things on which we have data,
- We'll use "observations" to refer to the measurements on each variable unit on each unit at a given point in time; so, in nearly every case, we will have multiple *observations* on each *unit*.

Notationally, I'll generally use i to index units, and t to index time, with the number of units usually being denoted by N and the number of time points by T (so that the total number of observations is NT). So,

- Y_{it} indicates a variable that varies over both units and time,
- $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ is the over-time mean of Y,
- $\bar{Y}_t = \frac{1}{N}\sum_{i=1}^{N} Y_{it}$ is the across-unit mean of Y, and
- $\bar{Y} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}$ is the grand mean of Y.

Some Conceptual Issues

Variation

The variation in a variable that has measurements both across units and over time has two possible corresponding dimensions of variation: *cross-sectional* and *temporal*. It is entirely possible to think of (and common to encounter) data that, despite being measured on both dimensions, only varies on one.

Example: Survey data from a tiny, mythical panel survey:

id	year	gender	pres	pid	approve
1	1998	female	clinton	dem	3
1	2000	female	clinton	dem	3
1	2002	female	bush	dem	5
1	2004	female	bush	dem	3
2	1998	male	clinton	gop	2
2	2000	male	clinton	gop	1
2	2002	male	bush	gop	4
2	2004	male	bush	gop	3
3	1998	male	clinton	gop	2
3	2000	male	clinton	gop	2
3	2002	male	bush	gop	4
3	2004	male	bush	dem	1

Note:

- Unsurprisingly, `gender` does not vary over time within a particular unit (respondent), while `president` does not vary across units at any given time point.
- `approve` – a five-point measure of presidential approval – varies over both time and units.
- `pid`, which measure party identification, does not vary over time for the first two units we see here, but does (albeit a tiny bit) for the third (and could do so for other observations in the data, in theory...).

A key point to remember is that variation is *information*...

- In the limit, a complete absence of variation one one dimension means that there is nothing one can say about that phenomenon on that dimension...

- E.g., if we only look at the first respondent in the above data, we can say nothing about party identification (either how it varies over time, or its effect on anything else in the data), since it is a constant.
- Likewise, if we only consider the data on 1998, we can't say anything about general differences across presidents, since that "variable" is constant in that year $\forall$ respondents.

- Similarly, any variable that varies only a little bit on one dimension can tell us comparatively little about phenomena on that dimension...
  - So, `pid` – which varies for only one of the three respondents here – will be of little help in explaining other within-respondent variables. In fact, anything we might try to say about the relationship between `pid` and (say) `gender` is going to be based on the *one observation* where some variation exists between the two.

All of this means that one needs to consider carefully "where" the variation in one's data is, and – more important – where one's theories suggest we should see variation as well. Consider a ridiculously simplistic rendition of two major theories of international relations:

- **Realism**
  - Variation at the *system* level...
  - *Change over time* is key (polarity, etc.), with
  - Little (cross–sectional) variation across states at a particular time.
- **Liberalism**
  - States are the key actors...
  - Suggests that cross-sectional variation will be important, as well as temporal change.

We'll discuss some ways of more formally considering cross-sectional and temporal variation in the coming few classes.

Aggregation

One thing variation across a dimension allows us to do is aggregate data along that dimension. Rather than considering variation along both dimensions, we can aggregate the data (e.g., consider means or other summary measures) along that dimension. With two dimensions of variation, we can aggregate either cross-sectionally or temporally. This can be useful, but can also cause problems and (occasionally) be somewhat misleading.

Consider cross-sectional aggregation first. For our mythical data, we'd have:

id	gender	pres	pid	approve
1	female	?	dem	3.50
2	male	?	gop	2.50
3	male	?	?	2.25

Note:

- Where there is no variation on that dimension, aggregation is both easy and loses no information; R1 is always female, so "collapsing" her four observations into one tells us just as much as the data above do.
- Where data vary on the dimension of aggregation, we have to choose a means of aggregating/combining those data. Two things here:
    - Sometimes this is not so hard. For `approve`, I just took means here. Note, however, that this loses some important information, including *any* ability to say anything about either change over time or temporally-varying covariates (such as `pres`).
    - For others, it isn't so easy, particularly for nominal/categorical covariates. How does one aggregate `pres`? Is it always equal to 0.5? Similarly, how to aggregate `pid`, when it varies over time? One could do something like the percentage of the time that the respondent identified with (e.g.) the GOP, but (again) that would lose important temporal variation.
- So, these data would be somewhat useful if we were solely interested in (say) the relationship between gender and presidential approval, and if we did not think that relationship was moderated by other factors (such as the identity or party of the sitting president). But, that seems pretty unlikely.

We can also aggregate by time period, which looks like this:

year	female	pres	pid	approve
1998	0.33	clinton	0.66(?)	2.33
2000	0.33	clinton	0.66(?)	2.00
2002	0.33	bush	0.66(?)	4.33
2004	0.33	bush	0.33(?)	2.33

- Once again, we can calculate means by year for a variable like `approve`; this will give us an idea of the trend in that variable over time.

- Likewise, variables that don't vary within time periods (`pres`) are easy to aggregate, and some categorical variables that do vary cross-sectionally (such as `gender`) can still be aggregated, albeit with a loss of information.
- Finally, some variables like `pid`, we can calculate an average (or other central tendency measure), but there is a substantial loss of information in doing so.

In both cases, aggregation can tempt one to commit the *ecological fallacy*: inferring individual-level relationships on the basis of aggregate data. Ecological inference can be done – it is a complicated topic not addressed in this class – but not through standard methods such as regression.

The punch line: *Aggregation (almost always) loses information.* While there are certainly instances where looking at higher levels of aggregation can tell us interesting things, it is almost never the case that we can't learn just as much about those things with disaggregated data; and it is almost always the case that disaggregated data can tell us things that aggregates cannot.

Data, Pooling, etc.

Terminology

We're going to be talking about data in which variables vary both over time and across cross-sectional units. We'll always refer to the units as $i = 1, 2, ...N$, and to the time points as $t = 1, 2, ...T$. The total number of *observations* (i.e., lines of data) is equal to NT. Some general conventions for naming these kind of data are:

- ***Panel data*** generally refers to data which are cross-sectionally dominated; that is, where N is significantly larger than T. Examples are the NES panel studies ($N = 2000, T = 3$) or the Panel Study of Income Dynamics (N = large, $T = 12$ or so). Such data usually have a fixed T, so that these data's asymptotics are in N, which is important (we'll come back to this).

- ***Time-series cross-sectional*** (**TSCS**) data usually means data in which either T is dominant, or $N \approx T$. These data are common in comparative politics. But, it can also refer to data where N is dominant, but T is larger than in panel data (e.g. all-dyads all-years IR data, with N = several thousand and $T = 50$ or more). Here, N is usually fixed, and the asymptotics are in T; moreover, if we have enough data, we can say something about the time-series properties of the data as well as the cross-sectional part.

- ***Repeated measures data*** is a term that gets used more in biostats. Its useful, because it can mean any of these things, but its also vague.

Data Structure

In panel or TSCS data, we have multiple lines of data for each unit of observation. Such data are arranged as follows:

id	t	Y	X_1	...
1	1	250	3.4	...
1	2	290	3.3	...
⋮	⋮	⋮	⋮	...
2	1	160	4.7	...
2	2	150	4.9	...
⋮	⋮	⋮	⋮	...

Its common to organize the data first according to units, and then within units by time period. (In fact, some statistical packages require this arrangement for certain tests to work properly, while others will resort your data in this way automatically).

Variation in the TSCS context

Variables in TSCS data can, obviously, vary across units, or over time, or both. Consider some data on Supreme Court justices, by year:

- A variable that measures whether or not the justice is from the `south` will vary only between justices, never "within" any particular justice.
- Conversely, a variable for the `party` of the sitting president will not vary across justices in any given year ("between"), but will vary over time ("within").
- And a variable reflecting whether or not the political party of the sitting president is the same as that of the president that appointed the justice in the first place will (potentially) vary both "between" and "within" a particular justice.

We can think of these three kinds of variation as reflecting variations around some mean. Consider, for example, the variable for the number of majority and dissenting opinions written by a justice in a given year (we call this variable `writing`). Simply examining the mean and standard deviation yields:

$$\mu = 17.94,\ \sigma = 14.14,\ \text{Minimum} = 0,\ \text{Maximum} = 103,\ NT = 1765$$

This variation occurs both "within" and "between" justices, however. One way to separate these two concepts is to consider the justice-specific mean $\bar{X}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} X_{it}$. This value represents the average within-justice level of writing; comparing these differences tells us

what the between-justice differences in writing levels are. The difference between this average and the observed value in any given year is $X_{it} - \bar{X}_i$; we think of this deviation as the within-justice variation around the mean.

If we examine the "between" versus the "within" variation in `writing` separately (an easy way is to use `Stata`'s `-xtsum-` command), we find:

$$\sigma_{\text{Between}} = 11.24 \text{ Min} = 0 \text{ Max} = 65.53$$
$$\sigma_{\text{Within}} = 8.46 \text{ Min} = \text{-}26.59 \text{ Max} = 85.24$$

This suggests that there is generally greater variation in levels of writing "between" justices than there is within any given justice's career. (This ought not be too surprising). We'll come back to these ideas again numerous times in the next few classes.

General TSCS Regression Issues

Think of a general regression model for cross-sectional data:

$$Y_i = \alpha + \beta X_i + u_i \tag{1}$$

This model assumes several things:

- All the usual OLS assumptions, plus
- that the constant term is constant across different is, and
- that the effect of any given variable X on Y is constant across observations (at least, to the extent that non-constancy isn't specified in the model, e.g., through interaction terms).

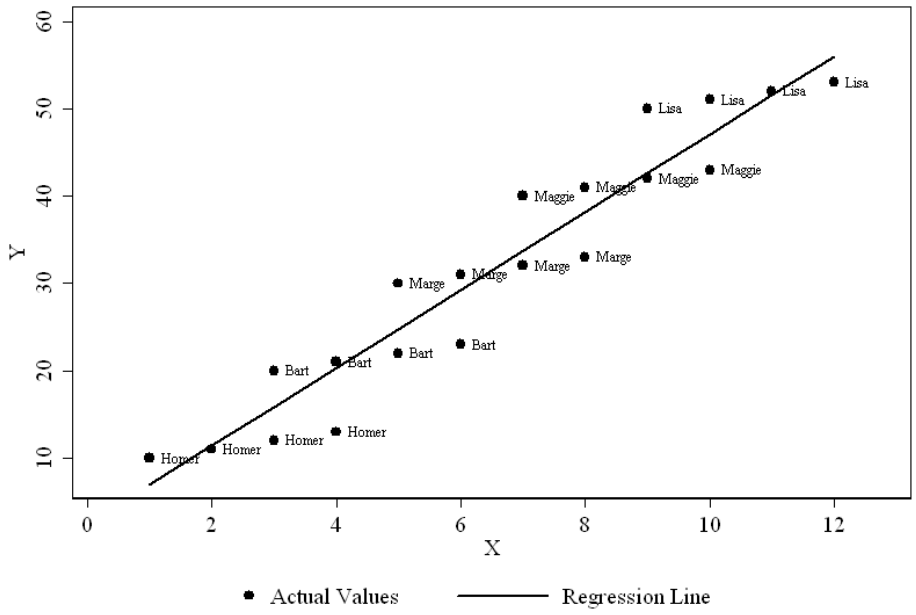
We can write a similar model in the TSCS context as follows:

$$Y_{it} = \alpha + \beta X_{it} + u_{it} \tag{2}$$

Note that this model assumes the same things as the earlier ones, especially about the effects of constants and covariates.

In *any* regression context, the two assumptions mentioned are critical; violating them leads to a form of specification bias. In the TSCS context, these two assumptions are often going to be problematic. This is because, since we're observing multiple units over time, there's often (in fact, usually) some reason to believe that there may be differences in either α or β over either i or t. Consider each of these possibilities.

Figure 1: Regression Results with Varying Intercepts



Variable Intercepts

One possible violation of the above assumptions is that the intercepts vary. The most common way this occurs is for different units to have varying intercepts:

$$Y_{it} = \alpha_i + \beta X_{it} + u_{it} \tag{3}$$

The slopes for each unit are the same, but the intercepts are different. Its also possible that the intercepts vary over time, rather than over units:

$$Y_{it} = \alpha_t + \beta X_{it} + u_{it} \tag{4}$$

or even over both i and t:

$$Y_{it} = \alpha_{it} + \beta X_{it} + u_{it} \tag{5}$$

Most of the time, however, it is unit differences that concern us most. If we have data that correspond to (3), but estimate a model like (2), we can get biased coefficients.

To see how this is true, consider the data in Figure 1. The actual slope (the effect of X on Y) is equal to 1.0; however we overestimate it significantly (here, we get $\hat{\beta} = 4.46$, with a standard error of 0.32) because of the different intercepts. Its just as likely that the bias

will be the other way, however; in most instances, we just don't know (since we likely don't know the actual values of the α_is).

Variable Slopes

The other obvious possibility is that we have a constant intercept, but the effects of X on Y differs across either units or (less likely) time; e.g.:

$$Y_{it} = \alpha + \beta_i X_{it} + u_{it} \tag{6}$$

We could also have variation in β over time, or even over both units and time.

A model like in (6) assumes that the regression lines all pass through the same point on the Y-axis, but that their slopes differ (perhaps tremendously). As a result, the estimate of $\hat{\beta}$ we'll get will be an "average" of those for the individual is. The idea of a common intercept, however, is a bit strange (at least to this social scientist), and is presented here as much for completeness as anything. More possible (/likely) is...

Variable Slopes and Intercepts

This is when things really start to get difficult. We might, for example, have variable slopes and intercepts for each unit i:

$$Y_{it} = \alpha_i + \beta_i X_{it} + u_{it} \tag{7}$$

Moreover, we could instead have different αs and βs for every time point, rather than for every unit:

$$Y_{it} = \alpha_t + \beta_t X_{it} + u_{it} \tag{8}$$

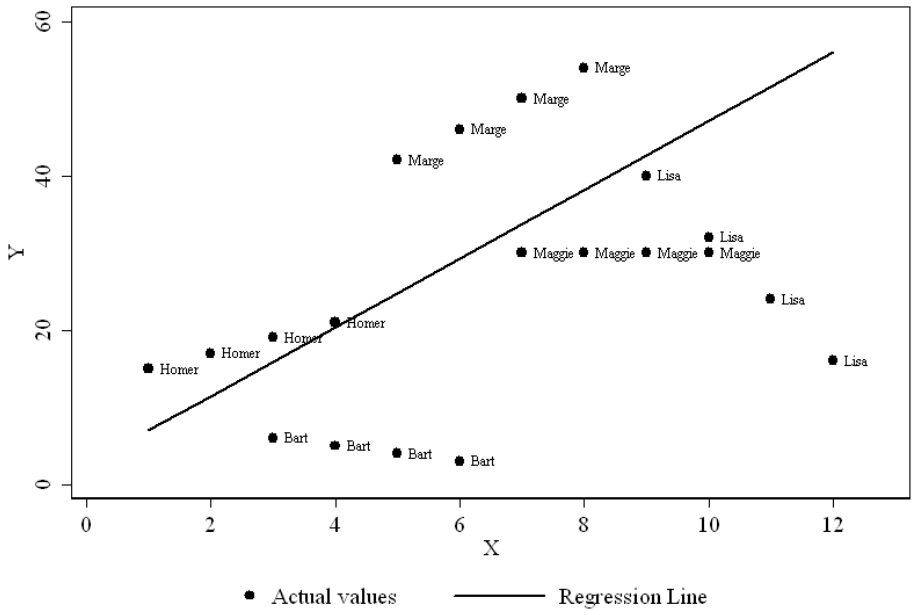
or for both different units and time points:

$$Y_{it} = \alpha_{it} + \beta_{it} X_{it} + u_{it} \tag{9}$$

The example in (7) is illustrated in Figure 2, which shows what you get if you estimate a model like (2) when the data correspond to (7).

Not surprisingly, the results you get are nonsensical, underestimating some slopes, overestimating others, and in some cases (e.g., Bart and Lisa) even getting the sign wrong. This points up how important accurately modeling slope- and intercept-variation in TSCS models can be.

Figure 2: Regression Results with Varying Slopes and Intercepts



The Error Term

Note, as well that, throughout all this discussion, we've been assuming that the error term u_{it} is (a) homoscedastic and (b) uncorrelated, both within and across i and t. Formally, that means we need to have:

$$u_{it} \sim \text{i.i.d.} N(0, \sigma^2) \; \forall \; i, t \tag{10}$$

If you stop and think about it, that's a pretty tall order. In particular, it requires that:

$$\begin{aligned} \text{Var}(u_{it}) &= \text{Var}(u_{jt}) \; \forall \; i \neq j \text{ (i.e., no cross-unit heteroscedasticity)} \\ \text{Var}(u_{it}) &= \text{Var}(u_{is}) \; \forall \; t \neq s \text{ (i.e., no temporal heteroscedasticity)} \\ \text{Cov}(u_{it}, u_{js}) &= 0 \; \forall \; i \neq j, \; \forall \; t \neq s \text{ (i.e., no auto- or spatial correlation)} \end{aligned}$$

Remember: Residuals are (among other things) just an indicator of how good a job the model does of explaining Y with $\mathbf{X}$. In that light, these assumptions are violated if (for example):

- Cross-unit differences mean that the model does a better job of explaining some units than others,

- Time effects (such as socialization, institutionalization, learning, or other such dynamics) cause the model to do a better or worse job of explaining Y over time,
- Omitted variables lead to residual correlation, either across units or (more commonly) over time.

While – in a linear model, at least – problems with the error term don't bias coefficient estimates, they can screw up one's inferences pretty badly. And in nonlinear models (logits and such) they can also lead to biases in the point estimates as well. We'll address these issues as we go along.

All this leads to the issue of...

Pooling

Pooling is nothing more than combining data, either across units or over time. Key to pooling is *exchangeability*: the notion that, conditional on the values of the covariates, any two observations within our data are considered to be the same ("exchangeable"). In the panel context, one aspect of this is what is known as "poolability".

Why Pool?

Several reasons:

- **Pooling adds data.** This is the number one reason for pooling data. If the assumption of poolability holds, we can get "better" (read: more precise) estimates of our $\hat{\beta}$s.
- **Generalizability**. Again, if a case can be made for model-conditional poolability, adding different cases means we can be more sure that our data generalize to broader sets of cases and/or longer time periods.

Issues in Pooling

Implicit in any panel data analysis, then, is the idea that the coefficients β do not vary over subsets of the data defined along N or T. In particular, in (say) the general, restrictive model (2):

$$Y_{it} = \alpha + \beta X_{it} + u_{it}$$

the implicit assumption is one of "exchangeability" – i.e., that all of the data come from the same "regime," that is,

- that the process governing the relationship between X and Y is exactly the same for each i,
- that the process governing the relationship between X and Y is the same for all t,

- that the process governing the us is the same $\forall i$ and t as well.

Bartels (1996) lays out a nice general discussion of pooling, outside the context of panel/TSCS data, but a lot of what he says is especially applicable to what we're going to be studying. Consider the general case of two "regimes":

$$Y_A = \beta'_A \mathbf{X}_A + u_A$$

$$Y_B = \beta'_B \mathbf{X}_B + u_B$$

based on N_A and N_B observations, respectively, where β_A and β_B are $K \times 1$ vectors of parameters including a constant term and where, for simplicity, the us meet all the usual requirements for regression errors, $\sigma^2_A = \sigma^2_B$, and $\text{Cov}(u_A, u_B) = 0$. The separate estimators are simply:

$$\hat{\beta}_{A,B} = (\mathbf{X}'_{A,B}\mathbf{X}_{A,B})^{-1}\mathbf{X}'_{A,B}Y_{A,B} \tag{11}$$

and the corresponding estimators of the variance-covariance matrices are:

$$\widehat{\text{Var}(\beta_{A,B})} = \hat{\sigma}^2_{A,B}(\mathbf{X}'_{A,B}\mathbf{X}_{A,B})^{-1} \tag{12}$$

Bartels shows that the pooled estimator $\hat{\beta_P}$ is then:

$$\begin{aligned} \hat{\beta}_P &= (\mathbf{X}'_A\mathbf{X}_A + \mathbf{X}'_B\mathbf{X}_B)^{-1}(\mathbf{X}'_A Y_A + \mathbf{X}'_B Y_B) \\ &= (\mathbf{X}'_A\mathbf{X}_A + \mathbf{X}'_B\mathbf{X}_B)^{-1}[\beta_A(\mathbf{X}'_A\mathbf{X}_A) + \beta_B(\mathbf{X}'_B\mathbf{X}_B)], \end{aligned} \tag{13}$$

which is a weighted combination of the two separate βs, the weights being inversely proportional to the variance-covariance matrix of that parameter vector.

In English: $\hat{\beta}_P$ is a combination of β_A and β_B, where the "better" / more precise of the two coefficients will dominate. Not surprisingly, this means that, *ceteris paribus*, the regime that will dominate $\hat{\beta}_P$ will be the one with:

- the larger N,
- the larger values of the coefficients, and/or
- the smaller standard errors of $\hat{\beta}$.

In addition, Bartels notes that the expectation of the pooled estimator $\hat{\beta_P}$ is:

$$\begin{aligned} \text{E}(\hat{\beta}_P) &= \beta_A + (\mathbf{X}'_A\mathbf{X}_A + \mathbf{X}'_B\mathbf{X}_B)^{-1}\mathbf{X}'_B\mathbf{X}_B(\beta_B - \beta_A) \\ &= \beta_B + (\mathbf{X}'_A\mathbf{X}_A + \mathbf{X}'_B\mathbf{X}_B)^{-1}\mathbf{X}'_A\mathbf{X}_A(\beta_A - \beta_B) \end{aligned} \tag{14}$$

This is a nice illustration of how pooling observations from different "regimes" can yield biases in coefficients (that is, if the latter term doesn't go to zero).

Testing

A standard way to assess whether or not this is the case is by comparing the "fit" of the separate regressions to that from the pooled model. Formally, we do this by considering the sums of squared residuals:

$$F = \frac{\frac{\hat{\mathbf{u}}_P'\hat{\mathbf{u}}_P - (\hat{\mathbf{u}}_A'\hat{\mathbf{u}}_A + \hat{\mathbf{u}}_B'\hat{\mathbf{u}}_B)}{K}}{\frac{(\hat{\mathbf{u}}_A'\hat{\mathbf{u}}_A + \hat{\mathbf{u}}_B'\hat{\mathbf{u}}_B)}{(N_A + N_B - 2K)}} \tag{15}$$

This test, which is distributed $F_{[K,(N_A+N_B-2K)]}$ is many things:

- Most generally, it is a test based on some number of linear restrictions of the original, general model in (9). As such, it can be formulated in a general way, as an F-test of restrictions on a standard OLS/GLS regression model.
- In the case where $K = 2$ (i.e., bivariate regression with a constant), this is equal to the t-test for $(\hat{\beta_A} - \hat{\beta_B})$.
- In a time-series setting, this is often referred to as a "Chow test" for structural stability in the parameter vector (e.g., when A and B represent two different time periods in the same time series).
- Finally, the test is only valid if $\sigma_A^2 = \sigma_B^2$; if not, then there's a more general test that is asymptotically correct (see, e.g., Davidson and MacKinnon 1993, but also the Stata FAQ on the topic).

Of course, if the pooling is simple (i.e., only a few covariates of interest, and a relatively small number of possible "groups"), another approach is to estimate a model with a multiplicative interaction term:

$$Y_{it} = \alpha + \beta X_{it} + \gamma B_{it} + \delta X_{it} B_{it} + u_{it}. \tag{16}$$

In this model, $\hat{\alpha}$ and $\hat{\beta}$ represent the intercepts and slopes, respectively, for the observations in group A, while $\hat{\alpha} + \hat{\gamma}$ and $\hat{\beta} + \hat{\delta}$ are the intercept and slope for the observations in group B (but, of course, you all knew that...). With this approach, we can simply test for the joint significance of $\hat{\gamma}$ and $\hat{\delta}$ to see whether groups A and B pool. Of course, this approach get very unwieldy if we have large numbers of variables in $\mathbf{X}$, and/or of there are many more than two groups; in those instances, the Chow/Wald test is a better approach.

Fractional Pooling

Bartels' (somewhat Bayesian) idea is to "fractionally pool" observations, based on some weight $\lambda \in [0, 1]$ which the researcher assigns to (say) subset A of the data. The result is a fractionally pooled estimator of β, call it $\hat{\beta}_\lambda$:

$$\hat{\beta}_\lambda = (\lambda^2 \mathbf{X}'_A \mathbf{X}_A + \mathbf{X}'_B \mathbf{X}_B)^{-1}(\lambda^2 \mathbf{X}'_A Y_A + \mathbf{X}'_B Y_B) \tag{17}$$

In this estimator,

- $\lambda = 0$ corresponds to the separate estimators for $\hat{\beta_A}$ and $\hat{\beta_B}$,
- $\lambda = 1$ is the "fully pooled" estimator in (13),
- $0 < \lambda < 1$ corresponds to a regression where data in regime A are given some "partial" weighting in their contribution towards an estimate of β.

Bartels lays out a sort of empirical-Bayes rationale for doing this; its a general technique, and pretty cool. On the other hand, I haven't seen it used that much, probably because many social scientists are nervous about the perceived arbitrariness of selecting (a) value(s) for λ (which, in turn, is probably a function of the fact that so few of us are Bayesians at heart...).

Pooling, Summarized

> "(R)oughly speaking, it makes sense to pool disparate observations if the underlying parameters governing those observations are sufficiently similar, but not otherwise."

That about says it. In terms of practical advice, remember the following:

- Exchangeability is something to be explored and tested, not assumed.
- With a small number of "groups" to be pooled across, and/or a relatively small set of covariates $\mathbf{X}$, estimating a model of the form in (16) and testing for the joint significance of the interactive terms is a viable option.
- For more complicated situations, a Chow/Wald test is a better alternative.

Analyzing Panel/TSCS Data in Stata

General Practices

When analyzing such data in Stata, its good practice to `-sort-` the data on the N and T identifier variables periodically.

The series of commands in Stata for analyzing such data all begin with the letters `-xt-` (for "**x**-sectional **t**ime-series"). We need to tell Stata that our data are of this format in order to use these commands. We do this by specifying the i and t variables:

```
. iis idvar

. tis timevar
```

Once we've done this, there are a number of regression-type commands that become available to us. Also, Stata has a few useful commands for managing TSCS/panel data...

The `-expand-` command

Stata's `-expand-` creates multiple "copies" of observations already in the data. This is good as a first step, when you have data that don't vary over time but are planing on adding/collecting some that does. Suppose we have a (very) small dataset on three countries – the U.S., the U.K., and Japan – which included data on variables that didn't vary over time (e.g., government type, etc.):

ID	x1	years	...	
US	250	7	...	
UK	290	9	...	
JP	150	5	...	
⋮	⋮	⋮	⋮	...

Suppose we wanted to collect ten years of data, 1991-2000, for each country (giving us $NT = 30$). To create a dataset with 30 lines of data, and with the existing X variables retained for each country, we would simply type:

```
. expand 10
```

This would give us a dataset that had 10 exact copies of each existing observation. We could then assign each line of data a year by typing:

```
. sort ID

. gen year = 1991

. quietly by ID : replace year=year[_n-1]+1 if year[_n-1]!=.
```

This gives us a dataset ready for inputting or merging time-varying data. The `-expand-` command will also take variables as an argument; so if, for example, you wanted to create a

number of years equal to the variable `years` in the data, you would type:

```
. expand years
```

which would create six copies of the observation for the U.S., eight of the U.K., and four for Japan.

The `-reshape-` command

Stata's `-reshape-` command is useful for converting data from "wide" to "long" format and back. Think of the way we usually arrange (e.g.) TSCS data as "long", in that we use rows rather than columns for storing information. So, if we had data on three years worth of GDP numbers for the three countries mentioned (i.e., $NT = 9$), we'd typically have it arranged as:

name	year	gdp
US	1980	280
US	1981	294
US	1982	303
UK	1980	121
UK	1981	124
UK	1982	131
JP	1980	176
JP	1981	192
JP	1982	212

Note that we could accomplish the same thing by storing the data with separate variables for each of the GDP-years:

name	gdp80	gdp81	gdp82
US	280	294	303
UK	121	124	131
JP	176	192	212

The `-reshape-` command converts data from one such format to the other. I won't go into the details of it right now (there are lots of options), but suffice it to say that, in many cases, you receive (e.g.) government data in "wide" format, and need to convert it to "long" format in order to analyze it. `-reshape-` makes this much easier.

The `-stack-` command

`-stack-` does exactly that; it takes existing variables and "stacks" them into a single column. This is useful when you have data that are in a variation of "wide" format, in that it can act as a combination of `-reshape-` and `-expand-`. Suppose your data look like this:

name	us	uk	jp
1980	280	121	176
1981	294	124	192
1982	303	131	212

The `-stack-` command will convert this onto normal ("long"-format) data in one fell swoop:

```
. stack us uk jp, into(gdp)
```

Your new data are now in "long" format, albeit minus labels (so keep good track of what goes where). `-stack-` is generally more useful for smaller datasets with few variables; otherwise, `-reshape-` is more flexible.

Pooling, tests, etc.

Stata makes testing for exchangeability and the like relatively easy. The `-test-` and `testparm` commands are very versatile in this regard (and, in fact, are a practical quantitative analysts best friend in a range of situations).

An Example: Left Governments and Unemployment in the OECD, 1978-1994

For an example, consider some data on average annual unemployment rates on 18 OECD countries for 17 years (1978–1994, inclusive). We're interested in seeing if – as theories might suggest – left governments generally seek to minimize unemployment (in exchange for higher inflation), while right governments minimize inflation (at the cost of higher unemployment). So, our lone X variable is the lagged percentage of the cabinet that are from left (i.e., labor, social democratic, or Green) parties. Moreover, we'll also try to discern if the dynamics of the unemployment/party control relationship are different in Anglo countries (that is, the UK and its former colonies) than in the rest of the OECD.

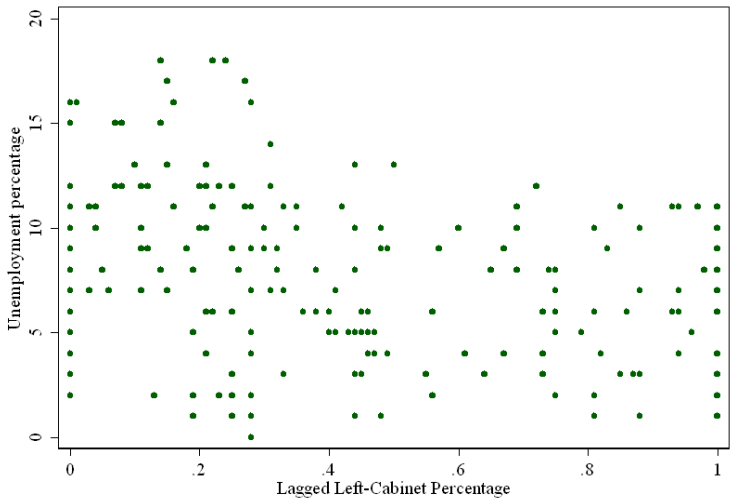
Here are what the data "look like," in terms of missing data and the like:

```
. xtdes

countryid:  1, 2, ..., 18                                          n =      18
    year:  78, 79, ..., 94                                         T =      17
           Delta(year) = 1; (94-78)+1 = 17
           (countryid*year uniquely identifies each observation)

Distribution of T_i:  min     5%    25%       50%       75%    95%    max
                       17     17     17        17        17     17     17

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-------------------
       18    100.00  100.00 |  1111111111111111
 ---------------------------+-------------------
       18    100.00         |  XXXXXXXXXXXXXXXXX
```

There are 18 countries, 17 years, and no missing data here. If we plot the two variables, the look like this:

```
.  twoway (scatter unemp leftcab, msymbol(smcircle))
```

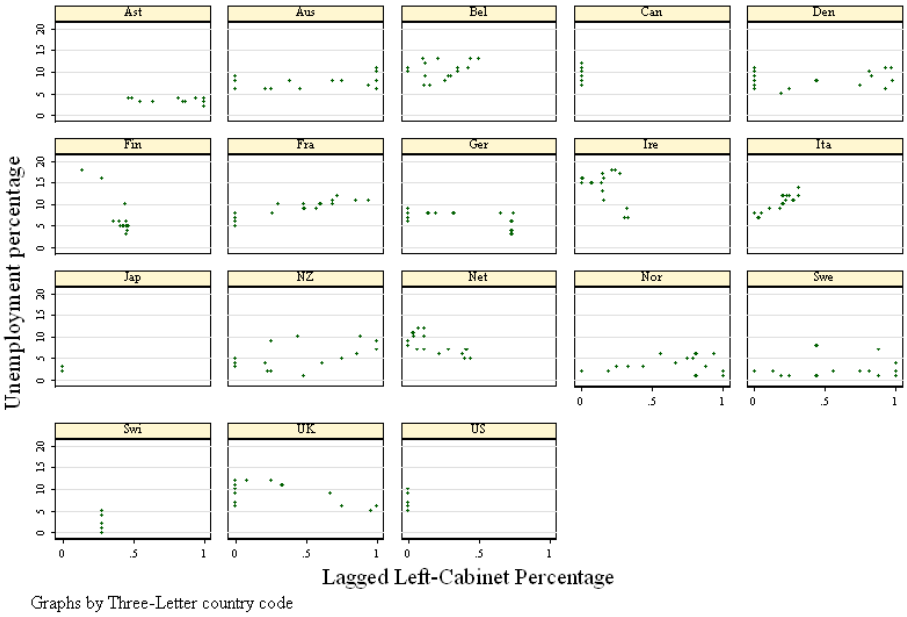
Figure 3: Unemployment and Left Cabinets, OECD 1978-1994



However, it's likely we're missing something here, a fact we can see if we examine the scatterplots country-by-country:

```
. twoway (scatter unemp leftcab, msymbol(smcircle)), by(country)
```

Figure 4: Unemployment and Left Cabinets, OECD 1978-1994, by Country



Note:

- Some countries have little (or no) variation on `leftcab`, and some have very little on `unemp` as well.
- Lots of different slopes, lots of different intercepts...

So, should we pool these data? Well, one way to find out is to test whether or not there are differences across important variables. Here, we chose `anglo`, a variable that is coded one for Australia, Canada, Ireland, New Zealand, the U.K., and the U.S., and zero elsewhere.

We'll start with a pooled regression:

```
. regress unemp leftcab

      Source |       SS       df       MS              Number of obs =     306
-------------+------------------------------           F(  1,   304) =   16.18
       Model |  238.221216     1  238.221216           Prob > F      =  0.0001
    Residual |  4475.05989   304  14.7205917           R-squared     =  0.0505
-------------+------------------------------           Adj R-squared =  0.0474
       Total |  4713.28111   305  15.4533807           Root MSE      =  3.8367

------------------------------------------------------------------------------
       unemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     leftcab |  -2.587552   .6432229    -4.02   0.000    -3.853285   -1.321819
       _cons |   7.722606   .3039998    25.40   0.000     7.124396    8.320816
------------------------------------------------------------------------------
```

Next, we can consider separate regressions:

```
. regress unemp leftcab if anglo==0

      Source |       SS       df       MS              Number of obs =     204
-------------+------------------------------           F(  1,   202) =    7.89
       Model |  111.011084     1  111.011084           Prob > F      =  0.0055
    Residual |  2841.98401   202  14.0692278           R-squared     =  0.0376
-------------+------------------------------           Adj R-squared =  0.0328
       Total |   2952.9951   203  14.5467739           Root MSE      =  3.7509

------------------------------------------------------------------------------
       unemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     leftcab |  -2.229366    .793658    -2.81   0.005    -3.794283   -.6644495
       _cons |   6.831329   .3969784    17.21   0.000     6.048576    7.614082
------------------------------------------------------------------------------

. regress unemp leftcab if anglo==1

      Source |       SS       df       MS              Number of obs =     102
-------------+------------------------------           F(  1,   100) =    1.82
       Model |  22.9381253     1  22.9381253           Prob > F      =  0.1808
    Residual |  1262.64035   100  12.6264035           R-squared     =  0.0178
-------------+------------------------------           Adj R-squared =  0.0080
```

```
       Total |  1285.57848   101  12.7284998           Root MSE      =  3.5534

------------------------------------------------------------------------------
       unemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     leftcab |  -1.391711   1.032547    -1.35   0.181    -3.440255    .6568333
       _cons |   8.959532   .4253936    21.06   0.000     8.115563      9.8035
------------------------------------------------------------------------------
```

This suggests that both the intercept and the slope are somewhat different between the two models. But, as long as they are estimated separately, we can't test for differences between them. To do so, we can estimate (and test on) an interactive model:

```
. gen angxleft=anglo*left

. regress unemp leftcab anglo angxleft

      Source |       SS       df       MS              Number of obs =     306
-------------+------------------------------           F(  3,   302) =   14.93
       Model |  608.656739     3   202.88558           Prob > F      =  0.0000
    Residual |  4104.62437   302  13.5914714           R-squared     =  0.1291
-------------+------------------------------           Adj R-squared =  0.1205
       Total |  4713.28111   305  15.4533807           Root MSE      =  3.6867

------------------------------------------------------------------------------
       unemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     leftcab |  -2.229366   .7800663    -2.86   0.005     -3.76442   -.6943128
       anglo |   2.128202   .5890937     3.61   0.000     .9689544    3.287451
    angxleft |   .8376555   1.325197     0.63   0.528    -1.770133    3.445444
       _cons |   6.831329     .39018    17.51   0.000     6.063513    7.599145
------------------------------------------------------------------------------

. testparm anglo angxleft

 ( 1)  anglo = 0
 ( 2)  angxleft = 0

       F(  2,   302) =   13.63
            Prob > F =    0.0000
```