POLS UG4792

Advanced Topics in Quantitative Research: Models for Panel and Time-Series Cross-Section Data

Gregory Wawro
Professor
Department of Political Science
Columbia University
420 W. 118th St.
New York, NY 10027

gjw10@columbia.edu

ACKNOWLEDGEMENTS

This course draws liberally on lecture notes prepared by Professors Neal Beck, Lucy Goodhart, George Jakubson, and Nolan McCarty. The course also draws from the following works:

Angrist, Joshua D. and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Baltagi, Badi H. 2008. *Econometric Analysis of Panel Data*, 4th edition. New York: John Wiley & Sons.

Beck, Nathaniel, and Jonathan N. Katz. 1995. "What To Do (and Not To Do) with Time-SeriesCross-Section Data in Comparative Politics." *American Political Science Review* 89:634-647.

Beck, Nathaniel, and Jonathan N. Katz. 2004. "Random Coefficient Models For Time-Series Cross-Section Data." Social Science Working Paper 1205. Division Of The Humanities And Social Sciences. California Institute Of Technology.

Beck, Nathaniel, Jonathan N. Katz, and Richard Tucker. 1998. "Taking Time Seriously: Time-SeriesCross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42:12601288.

Davidson, Russell and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics.* New York: Oxford University Press.

Greene, William H. 2011. *Econometric Analysis*, Seventh Edition, Pearson Prentice Hall.

Hsiao, Cheng. 2003. *Analysis of Panel Data.* 2nd ed. Cambridge: Cambridge University Press.

Wooldrige, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.

Zorn, Christopher J. W. 2001. "Generalized Estimating Equations Models for Correlated Data: A Review With Applications." *American Journal of Political Science* 45: 470–90.

Zorn, Christopher J. W. Notes for Advanced Maximum Likelihood, ICPSR Summer Program in Quantitative Methods (http://www.polisci.emory.edu/zorn/Classes/ICPSR2003/index.html).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Section 1

General Issues With Repeated Observations Data

1.1 Introduction

Definition: data in which we observe cross-sectional units at more than one time period.

- Repeated observations data (ROD) is a very valuable resource for finding empirical solutions to social science puzzles.
  - ➢ Quantity: increases samples sizes/provides more df.
  - ➢ Quality: enables us to answer questions in ways that cross-sectional data cannot.
- Example 1: Suppose we have a cross-sectional data set where we observe that the participation rate of Latinos in an election is 40%.
  - ➢ Could be the case that a given Latino voter has a 40% chance of participating in any given election.
  - ➢ Could also be the case that 40% of Latinos will vote in every election and 60% will never vote.
  - ➢ Can distinguish b/t these two scenarios if we get to observe these voters over time.

- Example 2: Democracy and Economic Growth. Some might argue that democratic governments cause higher economic growth (b/c of constraints on govt., etc.). Others might argue that the same conditions that enable populations to transition to and sustain democracies (e.g., respect for rule of law, sense of fairness) also lead them to have higher economic growth.

  ➢ By observing countries before and after transitions to democracy—while presumably accounting for (possibly unobservable) factors—we can potentially reject one of these hypotheses.

- Can estimate more elaborate models in terms of parameter variation (allow parameters to vary over individuals and/or over time).

- Standard distinction between types of ROD:

  1. T is large relative to N: Time-Series Cross-Section (TSCS) data.
  2. N is large relative to T: Panel data.

- But ROD also presents new challenges:

  ➢ Concerns about asymptotics—in N? in T? in both?

  ➢ How do we treat unit effects—as fixed or random?

  ➢ Complicated error structures.

  ➢ Issues of panel attrition.

- Consider the model

$$y_{it} = \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it}$$

  where

  - ➢ $i = 1,\ldots, N$ cross-sectional units (e.g., countries, states, households)
  - ➢ $t = 1,\ldots, T$ time units (e.g., years, quarters, months)
  - ➢ NT = total number of data points.

- Oftentimes, this model will be estimated pooling the data from different units and time periods.

- Advantages: can give more leverage in estimating parameters; consistent w/ "general" theories

1.2 Example 2 (cont.): Unobserved Country Effects and LSDV

- Consider

$$Growth_{it} = \beta_0 + \beta_1 Democracy_{it} + \beta_2 Z_{it} + \varepsilon_{it}$$

- We might be worried that our measure of democracy is actually proxying for other unmeasured factors specific to each country—i.e., there is cross-sectional heterogeneity in the data unaccounted for in the model.
- If these unit-specific factors are correlated with other variables in the model $\rightarrow$ omitted variable bias.
- Even if not, we will get larger standard errors because we are not incorporating sources of cross-country variation into the model.
- We could try to explicitly incorporate all the systematic factors that might lead to variation in growth, but places high demands in terms of data gathering.
- An alternative approach, which would be less demanding data-wise, is to explicitly include unit effects in the model:

$$Growth_{it} = \alpha_i + \beta_1 Democracy_{it} + \beta_2 Z_{it} + \varepsilon_{it}$$

- One way estimate this is to introduce a set of country dummies into the model:

$$\begin{aligned} Growth_{it} =& \beta_0 + \alpha_1 Country_1 + \ldots + \alpha_{N-1} Country_{N-1} \\ &+ \beta_1 Democracy_{it} + \beta_2 Z_{it} + \varepsilon_{it} \end{aligned}$$

  or

$$\begin{aligned} Growth_{it} =& \alpha_1 Country_1 + \ldots + \alpha_N Country_N \\ &+ \beta_1 Democracy_{it} + \beta_2 Z_{it} + \varepsilon_{it} \end{aligned}$$

  Need to avoid the dummy variable trap.

- We might also worry about time-specific effects (e.g., a general shock to the banking system hurts growth everywhere), which could introduce bias if they are not accounted for:

$$Growth_{it} = \alpha_i + \delta_t + \beta_1 Democracy_{it} + \beta_2 Z_{it} + \varepsilon_{it}$$

- To operationalize, we can introduce a $T - 1$ set of dummies for all but one time period.

- The degrees of freedom for the model are now $NT - k - N - (T - 1)$. The statistical significance of the country-specific and time-specific effects can be tested by using an F-test to see if the country/time dummies are jointly significant.

- The general approach of including unit-specific dummies is known as *Least Squares Dummy Variables* model, or *LSDV*.

1.3 Consequences of not accounting for heterogeneity

- If the α vary over individuals and we pool the data we can get bias in estimates of the slope and intercepts—see Figs. 1.1 and 1.2.

1.4 Testing for unit or time effects

- For LSDV (including an intercept), we want to test the linear hypothesis that

$$\alpha_1 = \alpha_2 = ... = \alpha_{N-1} = 0$$

- Can use an F-test:

$$F(N-1, NT-N-K) = \frac{(R^2_{UR} - R^2_R)/(N-1)}{(1-R^2_{UR})/(NT-N-K)}$$

  In this case, the unrestricted model is the one with the cross-sectional dummies (and hence different intercepts); the restricted model is the one with just a single intercept. A similar test could also be performed on the year dummies.

  ➢ Note that $N-1$ represents the number of new regressors in the unrestricted model and $NT-N-K$ represents the total number of data points minus the total number of parameters in the unrestricted model.

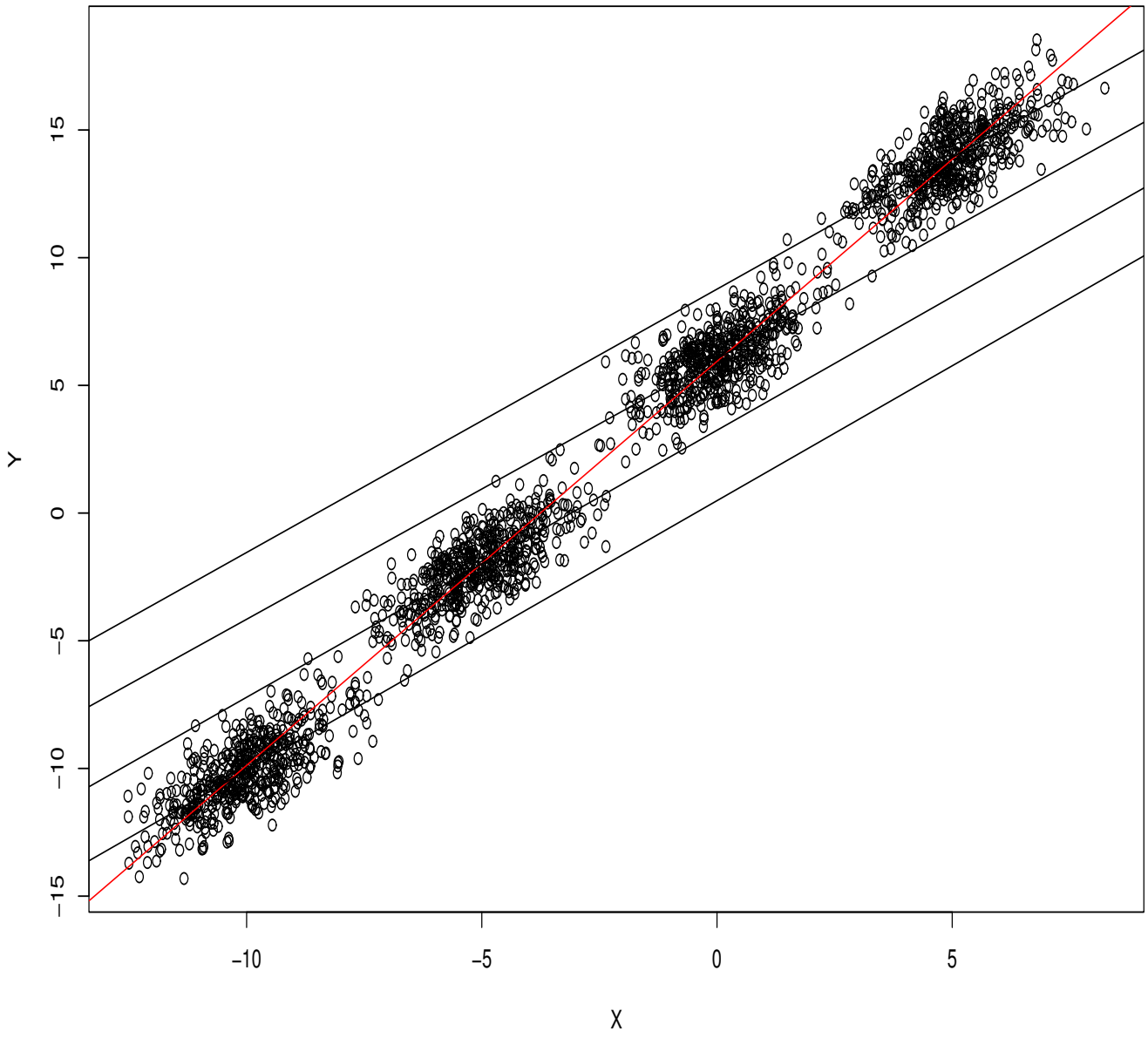


Figure 1.1: Heterogeneity Bias—overstating the effect

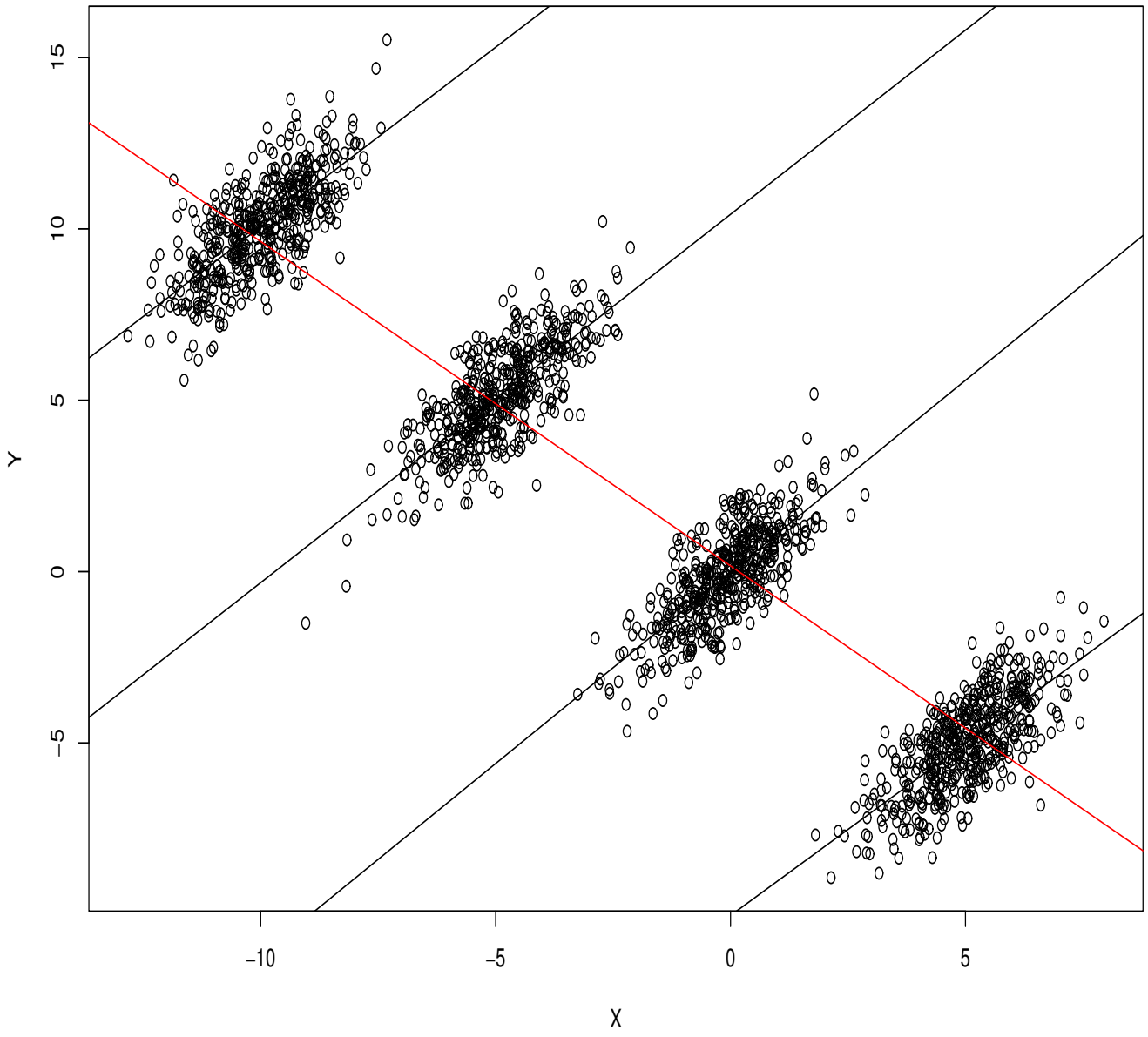


Figure 1.2: Heterogeneity Bias—incorrect sign

- To perform this test in Stata, after the `regress` command do:

1. If there are $N-1$ cross-sectional dummies and an intercept

   `test dummy1=dummy2=dummy3=dummy4=...=dummyN-1=0`

2. If there are N cross-sectional dummies and no intercept

   `test dummy1=dummy2=dummy3=dummy4=...=dummyN`

1.4.1 An alternative test

- Beck and Katz ('01 *IO*) argue that the Schwartz Criterion (SC) is superior to the standard F test for the presence of unit effects, because the SC imposes a higher penalty for including more explanatory variables.
- The SC provides a difficult test for the LSDV model where N is particularly large and separate dummies for each cross-sectional unit are specified.
- Assume a prior probability of the true model being K_1 and a prior conditional distribution of the parameters given that K_1 is the true model. Then choose the a posteriori most probable model.
- We choose the model that minimizes

$$SC = \ln(\mathbf{u}'\mathbf{u}/NT) + \frac{K \ln NT}{NT} \tag{1.1}$$

  where $\mathbf{u}$ is the NT vector of estimated residuals.
- Just choose the model that has the lowest SC.

Section 2

Matrix Algebra and OLS/GLS Review

2.1 Assumed knowledge

- Assume basic knowledge of vectors/matrices and terminology (transpose, idempotent, symmetric).
- Vector/matrix operations (addition, subtraction, multiplication, conformability, inverse, non-singular).

2.2 The cross-sectional regression model in matrix form

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{k1} + \cdots + \beta_K x_{K1} + \varepsilon_1 \\
&\vdots \\
y_i &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + \varepsilon_i \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{1n} + \cdots + \beta_k x_{kn} + \cdots + \beta_K x_{Kn} + \varepsilon_n
\end{aligned}
$$

- Let

$$
\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \\ \vdots \\ \beta_K \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{K1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1i} & \cdots & x_{Ki} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{Kn} \end{bmatrix}
$$

- Now the entire model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- For a specific observation:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

2.3 OLS using matrix notation

- The sum of squared residuals is

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}})^2$$

  where $\widehat{\boldsymbol{\beta}}$ is now a vector of unknown coefficients.

- The minimization problem:

$$\min_{\widehat{\boldsymbol{\beta}}} \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

- Expanding:

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}$$

  or

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - 2\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}$$

- The first order conditions now become:

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \widehat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{0}$$

- The solution then satisfies the least squares normal equations:

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

- If the inverse of $\mathbf{X}'\mathbf{X}$ exists (i.e., assuming full rank) then the solution is:

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- The variance-covariance matrix for the errors:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = E\left(\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_N\end{bmatrix} \times \begin{bmatrix}\varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_N\end{bmatrix}\right)$$

$$= \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_N) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \cdots & E(\varepsilon_2\varepsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ E(\varepsilon_N\varepsilon_1) & E(\varepsilon_N\varepsilon_2) & \cdots & E(\varepsilon_N^2) \end{bmatrix}$$

$$\text{var}(\varepsilon_i) = E[(\varepsilon_i - E(\varepsilon_i))^2] = E[\varepsilon_i^2]$$

and

$$\text{cov}(\varepsilon_i, \varepsilon_j) = E[(\varepsilon_i - E(\varepsilon_i))(\varepsilon_j - E(\varepsilon_j))] = E(\varepsilon_i\varepsilon_j)$$

- If no heteroskedasticity and no serial correlation, then

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}_n = \begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ \vdots & & \ddots & \\ 0 & 0 & ... & \sigma^2 \end{bmatrix}$$

- We use this to find the variance-covariance matrix for $\widehat{\boldsymbol{\beta}}_{OLS}$:

$$\text{var}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = E\left[\left(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}\right)'\right]$$

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\end{aligned}$$

- Thus:

$$\begin{aligned}\text{var}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) &= E[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})((\mathbf{X}'\mathbf{X}^{-1})\mathbf{X}'\boldsymbol{\varepsilon})'] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\end{aligned}$$

- Using a theorem on the decomposition of the variance and passing through the expectations operator gives

$$\text{var}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- If the errors are nonspherical, then we have more complicated variance-covariance matrix structures.
- For heteroskedasticity:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{\Omega} = \begin{bmatrix}\sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & \sigma_n^2\end{bmatrix}$$

- For serial correlation—of the AR(1) variety:

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\boldsymbol{\Omega} = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^{T-3} \\ \vdots & & & & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & & \dots & 1 \end{bmatrix}$$

- With non-spherical errors, GLS is preferred to OLS, since the former and not the latter is BLUE:

$$\widehat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y})$$

  and

$$\text{var}\left(\widehat{\boldsymbol{\beta}}_{GLS}\right) = \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}.$$

- Of course, $\boldsymbol{\Omega}$ contains population parameters that we need to estimate, which means we are doing FGLS instead of GLS.

2.4 Matrix notation for panel/TSCS data

$$\underset{1\times 1}{y_{it}} \qquad \underset{T\times 1}{\mathbf{y}_i} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix} \qquad \underset{NT\times 1}{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1T} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2T} \\ \vdots \\ y_{N1} \\ y_{N2} \\ \vdots \\ y_{NT} \end{bmatrix}$$

$$\underset{K\times 1}{\mathbf{x}_{it}} = \begin{bmatrix} x_{1it} \\ x_{2it} \\ \vdots \\ x_{Kit} \end{bmatrix} \qquad \underset{T\times K}{\mathbf{X}_i} = \begin{bmatrix} x_{1i1} & x_{2i1} & \dots & x_{Ki1} \\ x_{1i2} & x_{2i2} & \dots & x_{Ki2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1iT} & x_{2iT} & \cdots & x_{KiT} \end{bmatrix} \qquad \underset{NT\times K}{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}$$

$$\underset{T\times 1}{\boldsymbol{\iota}} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \underset{T\times T}{\mathbf{I}_T} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

$$\underset{k\times l}{\mathbf{A}} \otimes \underset{m\times n}{\mathbf{B}} = \underset{km\times ln}{\begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1l}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2l}\mathbf{B} \\ \vdots & & \ddots & \vdots \\ a_{k1}\mathbf{B} & a_{k2}\mathbf{B} & \dots & a_{kl}\mathbf{B} \end{bmatrix}}$$

$$\underset{n\times n}{\mathbf{I}_N} \otimes \underset{k\times k}{\mathbf{V}} = \underset{nk\times nk}{\begin{bmatrix} \mathbf{V} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{V} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{V} \end{bmatrix}}$$

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t = \frac{1}{T}\boldsymbol{\iota}'\mathbf{x} \qquad \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it} = \frac{1}{T}\boldsymbol{\iota}'\mathbf{x}_i$$

$$\sum_{i=1}^{n} x_i y_i = \mathbf{x}'\mathbf{y}$$

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'$$

Section 3

Fixed Effects Estimators

3.1 LSDV as Fixed Effects

- Least squares dummy variable estimation is also known as **fixed effects**, because it assumes that the unobserved effect for a given cross-sectional unit or time period can be estimated as a given, *fixed* effect.
- Can think of this as fixed in repeated samples (e.g., France is France) as opposed to randomly drawn.
- Let the original model be

$$y_{it} = \alpha_i^* + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it} \tag{3.1}$$

- We can rewrite this in vector form as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{\iota} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \alpha_1^* + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\iota} \\ \vdots \\ \mathbf{0} \end{bmatrix} \alpha_2^* + \ldots + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \boldsymbol{\iota} \end{bmatrix} \alpha_N^* + \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_N \end{bmatrix} \tag{3.2}$$

where

$$\underset{T\times 1}{\mathbf{y}_i} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}, \quad \underset{T\times K}{\mathbf{X}_i} = \begin{bmatrix} x_{1i1} & x_{2i1} & \ldots & x_{Ki1} \\ x_{1i2} & x_{2i2} & \ldots & x_{Ki2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1iT} & x_{2iT} & \ldots & x_{KiT} \end{bmatrix}, \quad \underset{T\times 1}{\boldsymbol{\iota}} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

$$\underset{T\times 1}{\mathbf{u}_i} = \begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{bmatrix},$$

$$E[\mathbf{u}_i] = \mathbf{0}, \quad E[\mathbf{u}_i\mathbf{u}_i'] = \sigma_u^2\mathbf{I}_T, \quad E[\mathbf{u}_i\mathbf{u}_j'] = \mathbf{0} \text{ if } i \neq j.$$

- These assumptions regarding u_{it} mean that the OLS estimator for eq. 3.2 is BLUE.

- To obtain the OLS estimators of α_i^* and $\boldsymbol{\beta}$, we minimize:

$$S = \sum_{i=1}^{N} \mathbf{u}_i'\mathbf{u}_i = \sum_{i=1}^{N} (\mathbf{y}_i - \boldsymbol{\iota}\alpha_i^* - \mathbf{X}_i\boldsymbol{\beta})'(\mathbf{y}_i - \boldsymbol{\iota}\alpha_i^* - \mathbf{X}_i\boldsymbol{\beta}).$$

- Take partial derivatives wrt to α_i^*, set equal to zero and solve to get:

$$\hat{\alpha}_i^* = \bar{y}_i - \boldsymbol{\beta}'\bar{\mathbf{x}}_i \tag{3.3}$$

  where

$$\bar{y}_i = \sum_{t=1}^{T} y_{it}/T, \quad \bar{\mathbf{x}}_i = \sum_{t=1}^{T} \mathbf{x}_{it}/T.$$

- Substitute our estimate for $\hat{\alpha}_i^*$ in S, take partial derivatives wrt $\boldsymbol{\beta}$, set equal to zero and solve:

$$\hat{\boldsymbol{\beta}}_{\text{CV}} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\right]^{-1} \left[\sum_{i=1}^{N}\sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(y_{it} - \bar{y}_i)\right]$$

- Including separate dummies for each cross-sectional unit will produce estimates of the unit-specific effects.

- While this may be desirable, it does come at some cost—possibly inverting a large matrix of 0s and 1s.

- Another way to compute this estimator w/o including dummies is to subtract off the time means:

$$\bar{y}_i = \alpha_i^* + \boldsymbol{\beta}'\bar{\mathbf{x}}_i + \bar{u}_i \tag{3.4}$$

- If we estimated $\boldsymbol{\beta}$ in this equation by OLS (constraining $\alpha_i^* = \alpha^* \; \forall \; i$), it will produce what is known as the "Between Effects" estimator, or $\boldsymbol{\beta}_{BE}$, which shows how the mean level of the dependent variable for each cross-sectional unit varies with the mean level of the independent variables.

- Subtracting eq. 3.4 from eq. 3.1 gives

$$(y_{it} - \bar{y}_i) = (\alpha_i^* - \alpha_i^*) + \boldsymbol{\beta}'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) + (u_{it} - \bar{u}_i)$$

  or

$$(y_{it} - \bar{y}_i) = \boldsymbol{\beta}'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) + (u_{it} - \bar{u}_i)$$

- Running OLS on this equation gives results identical to LSDV.

- Sometimes called the *within-group estimator*, because it uses only the variation in y_{it} and $\mathbf{x}_{it}$ within each cross-sectional unit to estimate the $\boldsymbol{\beta}$ coefficients.

- Any variation *between* cross-sectional units is assumed to spring from the unobserved fixed effects.

- Another way to approach this is to pre-multiply each cross-sectional unit equation ($\mathbf{y}_i = \boldsymbol{\iota}\alpha_i^* + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i$) by a $T \times T$ idempotent "sweep" matrix:

$$\mathbf{Q} = \mathbf{I}_T - \frac{1}{T}\boldsymbol{\iota}\boldsymbol{\iota}'$$

- This has the effect of sweeping out the α_i^*s and transforming the variables so that the values for each individual are measured in terms of deviations from their means over time:

$$\mathbf{Q}\mathbf{y}_i = \mathbf{Q}\boldsymbol{\iota}\alpha_i^* + \mathbf{Q}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}\mathbf{u}_i \tag{3.5}$$

$$= \mathbf{Q}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}\mathbf{u}_i \tag{3.6}$$

- Running OLS on this regression gives

$$\hat{\boldsymbol{\beta}}_{\text{CV}} = \left[\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{Q}\mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{Q}\mathbf{y}_i\right]$$

- The variance-covariance matrix is

$$\text{var}[\hat{\boldsymbol{\beta}}_{\text{CV}}] = \sigma_u^2 \left[\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{Q}\mathbf{X}_i\right]^{-1}$$

- We can compute an estimate of σ_u^2 as

$$\hat{\sigma}_u^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(NT - N - k)$$

where

$$\hat{\mathbf{u}}_i = \mathbf{Q}\mathbf{y}_i - \mathbf{Q}\mathbf{X}_i\hat{\boldsymbol{\beta}}_{\text{CV}}$$

- Properties of $\boldsymbol{\beta}_{\text{CV}}$: unbiased and consistent whether N or T or both tend to infinity.
- Note that the OLS estimate of α_i^* is unbiased, but is consistent only as $T \to \infty$.

  ➢ With LSDV consistency is an issue: **incidental parameters problem**.
- Key advantage of FE estimators: can have correlation between $\mathbf{x}_{it}$ and α_i^*.
- A key drawback: if time-invariant regressors are included in the model, the standard FE estimator will not produce estimates for the coefficients on these variables (perfect collinearity in LSDV).

  ➢ There is an IV approach to produce estimates, but requires some exogeneity assumptions that are difficult to meet in practice.
- The effects of slow-moving variables can also be estimated very imprecisely due to collinearity.

3.2 Application: Economic growth in 14 OECD countries

- Garrett examines the political component of economic performance in his '98 book *Partisan Politics in the Global Economy.*
- Question: how does labor centralization and left control of the government affect economic growth (esp. in terms of an interaction effect)?
  - ➢ Tests the social democratic corporatist model of economic performance: congruence b/t partisan political control and labor bargaining enhances performance.
  - ➢ I.e., centralized labor bargaining goes well w/ politically powerful left parties; decentralized bargaining is more congruent when right parties are in power.
- Data: 14 OECD countries observed annually from 1966–1990 (i.e., $N = 14; T = 25$).
- Dependent variable: GDP.
- Explanatory vars:
  - ➢ Oil dependence (OIL),
  - ➢ Overall OECD GDP growth, weighted for each country by its trade with the other OECD nations (DEMAND),
  - ➢ proportion of cabinet posts occupied by left parties (LEFTLAB),
  - ➢ degree of centralized labor bargaining as a measure of corporatism (CORP),
  - ➢ interaction b/t CORP and LEFTLAB (CLINT).

- OLS on this equation gives:

```
. regress gdp oild demand corp leftlab clint

      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F(  5,   344) =   11.30
       Model |  291.283034     5  58.2566069           Prob > F      =  0.0000
    Residual |  1773.90603   344  5.15670357           R-squared     =  0.1410
-------------+------------------------------           Adj R-squared =  0.1286
       Total |  2065.18906   349  5.91744717           Root MSE      =  2.2708

------------------------------------------------------------------------------
         gdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        oild |   -15.2321   4.572497    -3.33   0.001    -24.22567   -6.238529
      demand |   .0049977    .000999     5.00   0.000     .0030328    .0069625
        corp |  -1.139716   .3043989    -3.74   0.000    -1.738433   -.5409982
     leftlab |  -1.483549   .3844653    -3.86   0.000    -2.239747   -.7273499
       clint |   .4547183   .1233779     3.69   0.000     .2120482    .6973883
       _cons |   5.919865   .7356383     8.05   0.000      4.47295     7.36678
------------------------------------------------------------------------------
```

- Including dummies for each country (except one) gives:

```
. regress gdp oild demand corp leftlab clint Icc_2 Icc_3 Icc_4 Icc_5 Icc_6 Icc_
> 7 Icc_8 Icc_9 Icc_10 Icc_11 Icc_12 Icc_13 Icc_14;

      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F( 18,   331) =    9.16
       Model |  686.921905    18  38.1623281           Prob > F      =  0.0000
    Residual |  1378.26716   331  4.16394912           R-squared     =  0.3326
-------------+------------------------------           Adj R-squared =  0.2963
       Total |  2065.18906   349  5.91744717           Root MSE      =  2.0406

------------------------------------------------------------------------------
         gdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        oild |  -25.59808   5.946569    -4.30   0.000    -37.29592   -13.90025
      demand |   .0084949    .001129     7.52   0.000      .006274    .0107158
        corp |  -.2500641   .6654194    -0.38   0.707    -1.559048     1.05892
     leftlab |  -1.172257   .4468775    -2.62   0.009    -2.051335   -.2931785
       clint |   .5030912   .1596682     3.15   0.002     .1889988    .8171836
       Icc_2 |  -.4136903   .5988751    -0.69   0.490    -1.591772    .7643909
       Icc_3 |  -2.090873   .7898403    -2.65   0.009    -3.644613   -.5371336
       Icc_4 |  -2.159732   .7633224    -2.83   0.005    -3.661307    -.658157
       Icc_5 |  -2.587796   1.091167    -2.37   0.018    -4.734293   -.4412985
       Icc_6 |   .6289216   .8440104     0.75   0.457    -1.031379    2.289222
       Icc_7 |  -1.796217   1.255454    -1.43   0.153    -4.265892    .6734584
       Icc_8 |  -3.993015   1.891938    -2.11   0.036    -7.714754   -.2712759
       Icc_9 |  -.8709414    1.02877    -0.85   0.398    -2.894693     1.15281
      Icc_10 |  -1.449112   1.301281    -1.11   0.266    -4.008935    1.110711
      Icc_11 |  -3.893792   1.607724    -2.42   0.016    -7.056438   -.7311463
      Icc_12 |  -3.489515    1.29017    -2.70   0.007    -6.027481   -.9515491
      Icc_13 |   -3.10808   1.477907    -2.10   0.036    -6.015355    -.200806
      Icc_14 |   2.929627   .6076861     4.82   0.000     1.734213    4.125041
       _cons |   3.374094   1.365808     2.47   0.014     .6873365    6.060852
------------------------------------------------------------------------------
```

- The F-test to determine if the dummies should be included gives:

```
. test Icc_2=Icc_3=Icc_4=Icc_5=Icc_6=Icc_7=Icc_8=Icc_9=Icc_10=Icc_11=Icc_12=Icc
> _13=Icc_14=0;

       F( 13,   331) =    7.31
            Prob > F =    0.0000
```

- Finally, estimating the fixed effects model:

```
. xtreg gdp oild demand corp leftlab clint, fe ;

Fixed-effects (within) regression               Number of obs      =       350
Group variable (i): country                     Number of groups   =        14

R-sq:  within  = 0.2315                         Obs per group: min =        25
       between = 0.0461                                        avg =      25.0
       overall = 0.0424                                        max =        25

                                                F(5,331)           =     19.94
corr(u_i, Xb)  = -0.7104                        Prob > F           =    0.0000

------------------------------------------------------------------------------
         gdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        oild |  -25.59808   5.946569    -4.30   0.000    -37.29592   -13.90025
      demand |   .0084949    .001129     7.52   0.000      .006274    .0107158
        corp |  -.2500641   .6654194    -0.38   0.707    -1.559048     1.05892
     leftlab |  -1.172257   .4468775    -2.62   0.009    -2.051335   -.2931785
       clint |   .5030912   .1596682     3.15   0.002     .1889988    .8171836
       _cons |    1.78165   1.961666     0.91   0.364    -2.077255    5.640556
-------------+----------------------------------------------------------------
     sigma_u |  1.9296773
     sigma_e |  2.0405757
         rho |  .47208946   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(13, 331) =     7.31                 Prob > F = 0.0000
```

3.3 Differences in Differences Estimation

- **Differences-in-differences (DD)** has gained in popularity as a means for investigating causal relationships.
- Especially popular for policy evaluation.
- Essentially a version of the fixed effects estimator using aggregate data.
- Let A indicate the control group, B denote the treatment group, $dB = 1$ if an observation is in B ($= 0$ otherwise), and $d2 = 1$ is an observation occurs in period 2 (e.g., post-policy change).
- Then we can analyze the impact of a policy change using

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u \tag{3.7}$$

  - ➢ dB: accounts for differences between A and B prior to the policy change.
  - ➢ $d2$: captures differences in outcome in the absence of a policy change.
  - ➢ δ_1: coefficient of interest.

- Interesting interpretation for OLS estimate of δ_1 as a differences-in-differences estimator:

$$\hat{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1}), \tag{3.8}$$

  where $\bar{y}_{A,1}$ denotes the sample average for the control group in period 1, $\bar{y}_{A,2}$ is the sample average for that group in period 2, etc.

- This accounts for both group-specific and time-specific effects and is superior to looking at just $(\bar{y}_{B,2} - \bar{y}_{B,1})$ (changes in y could be due to some other factor changing over time that could affect both A and B) or $(\bar{y}_{B,2} - \bar{y}_{A,2})$ (changes in y could be due to other differences b/t treatment and control groups).

- Note: point estimates and standard errors obtained from OLS on eq. 3.7 have desirable properties assuming well-behaved disturbances.

- Generalizations:

  ➢ Can add more covariates to eq. 3.7 to account for other factors that might affect outcome; interpretation of δ_1 is essentially unchanged.

  ➢ Can use non-binary/continuous treatment variables.

  ➢ Can have more than two time periods—e.g., a policy indicator replaces the $d2 \cdot dB$ interaction, defining for what groups and time periods the policy is in effect.

  ➢ Can have a full set of dummies for all groups, time periods, and for all pairwise interactions.

➢ Can further refine the definition of treatment & control groups:

$$y = \beta_0 + \beta_1 dB + \beta_2 dC + \beta_3 dB \cdot dC + \delta_0 d2 + \delta_1 d2 \cdot dB + \delta_2 d2 \cdot dC + \delta_3 d2 \cdot dB \cdot dC + u$$

where C denotes a subgroup within the treatment group B (say, an age category): **differences-in-differences-in-differences**.

➢ Now the main coefficient of interest is

$$\hat{\delta}_3 = [(\bar{y}_{B,C,2} - \bar{y}_{B,C,1}) - (\bar{y}_{B,\sim C,2} - \bar{y}_{B,\sim C,1})] - [(\bar{y}_{A,C,2} - \bar{y}_{A,C,1}) - (\bar{y}_{A,\sim C,2} - \bar{y}_{A,\sim C,1})]$$

- Potential pitfall: it is possible that composition of the treatment and control group changes as a result of treatment.

  ➢ E.g., in an analysis of the impact of welfare programs on incentive to work, need to worry about poor people (who might have weak labor force attachment) moving to states with more generous social welfare benefits.

Section 4

Random Effects Estimators

4.1 Introduction

- Fixed effects is completely appropriate if we believe that the unit-specific effects are indeed fixed, estimable amounts that we can calculate for each cross-sectional observation.
- Thus, we believe that Sweden will always have an intercept of 1.2 units (for instance). If we were able to take another sample, we would once again estimate the same intercept for Sweden. There are cases, however, where we may not believe that we can estimate some fixed amount for each country.
- In particular, assume that we have a panel data model run on 20 countries, but which should be generalizable to 100 different countries. We cannot estimate the given intercept for each country or each type of country because we don't have all of them in the sample for which we estimate the model.
- In this case, we might want to estimate the $\boldsymbol{\beta}$s on the explanatory variables taking into account that there could be country-specific effects that would enter as random shocks from a known distribution.
- If we go this route, we will be estimating a **random effects model**.

4.2 Deriving the random effects estimator

- Also called variance components b/c we can set up the disturbance in this way:

$$v_{it} = \alpha_i + \lambda_t + u_{it}.$$

- We make the following crucial assumptions about the variance and covariances of these components:

$$E[\alpha_i] = E[\lambda_t] = E[u_{it}] = 0$$

$$E[\alpha_i \lambda_t] = E[\alpha_i u_{it}] = E[\lambda_t u_{it}] = 0$$

$$E[\alpha_i \alpha_j] = \begin{cases} \sigma_\alpha^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$E[\lambda_t \lambda_s] = \begin{cases} \sigma_\lambda^2 & \text{if } t = s \\ 0 & \text{if } t \neq s \end{cases}$$

$$E[u_{it} u_{js}] = \begin{cases} \sigma_u^2 & \text{if } i = j, t = s \\ 0 & \text{otherwise} \end{cases}$$

$$E[\alpha_i \mathbf{x}_{it}'] = E[\lambda_t \mathbf{x}_{it}'] = E[u_{it} \mathbf{x}_{it}'] = \mathbf{0}$$

- Note that the $\text{var}[y_{it}|\mathbf{x}_{it}] = \sigma_y^2 = \sigma_\alpha^2 + \sigma_\lambda^2 + \sigma_u^2$.

- Let's add a general intercept to our model and set $\lambda_t = 0 \; \forall \; t$:

$$y_{it} = \mu + \boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i + u_{it} \tag{4.1}$$

- We can rewrite this in vector form:

$$\mathbf{y}_i = \tilde{\mathbf{X}}_i \boldsymbol{\delta} + \mathbf{v}_i \tag{4.2}$$

where

$$\underset{T\times(K+1)}{\tilde{\mathbf{X}}_i} = \begin{bmatrix} \boldsymbol{\iota} & \mathbf{X}_i \end{bmatrix}, \quad \underset{(K+1)\times 1}{\boldsymbol{\delta}} = \begin{bmatrix} \mu \\ \boldsymbol{\beta} \end{bmatrix}, \quad \underset{T\times 1}{\mathbf{v}_i} = \begin{bmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{iT} \end{bmatrix}, \quad v_{it} = \alpha_i + u_{it}$$

- The variance-covariance matrix of the T disturbance terms $\mathbf{v}_i$ is:

$$\mathbf{V} = E[\mathbf{v}_i \mathbf{v}_i'] = \begin{bmatrix} (\sigma_u^2 + \sigma_\alpha^2) & \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & (\sigma_u^2 + \sigma_\alpha^2) & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \vdots & & & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & (\sigma_u^2 + \sigma_\alpha^2) \end{bmatrix}$$

$$= \sigma_u^2 \mathbf{I}_T + \sigma_\alpha^2 \boldsymbol{\iota}\boldsymbol{\iota}'$$

- Note that

$$\mathbf{V}^{-1} = \frac{1}{\sigma_u^2} \left[\mathbf{I}_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2} \boldsymbol{\iota}\boldsymbol{\iota}' \right].$$

- The full variance-covariance matrix for all the NT observations is:

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{V} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{V} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{V} \end{bmatrix} = \mathbf{I}_N \otimes \mathbf{V}$$

- To produce parameter estimates, we could proceed in the manner that we did with the FE estimator—i.e., pre-multiply by $\mathbf{Q}$ and run OLS.
  - ➢ This will give unbiased and consistent estimates.
  - ➢ However, if the α_i are assumed to be random rather than fixed, the CV estimator is not BLUE; the GLS estimator is.
  - ➢ Intuition: v_{it} and v_{is} both contain α_i, inducing correlation in the disturbances; suggests that GLS would be appropriate to get efficient estimates; ignoring info about the DGP.

4.3 GLS Estimation

- The normal equations for the GLS estimator are

$$\left[\sum_{i=1}^{N} \tilde{\mathbf{X}}_i' \mathbf{V}^{-1} \tilde{\mathbf{X}}_i\right] \hat{\boldsymbol{\delta}}_{\text{GLS}} = \sum_{i=1}^{N} \tilde{\mathbf{X}}_i' \mathbf{V}^{-1} \mathbf{y}_i$$

- We could write the GLS estimator simply as

$$\hat{\boldsymbol{\delta}}_{\text{GLS}} = \left[\sum_{i=1}^{N} \tilde{\mathbf{X}}_i' \mathbf{V}^{-1} \tilde{\mathbf{X}}_i\right]^{-1} \sum_{i=1}^{N} \tilde{\mathbf{X}}_i' \mathbf{V}^{-1} \mathbf{y}_i \tag{4.3}$$

- But let's unpack this to get a better understanding of what is going on w/ this estimator. We can rewrite the inverse of the variance-covariance matrix as

$$\mathbf{V}^{-1} = \frac{1}{\sigma_u^2}\left[\mathbf{I}_T - \frac{1}{T}\boldsymbol{\iota}\boldsymbol{\iota}' + \psi \cdot \frac{1}{T}\boldsymbol{\iota}\boldsymbol{\iota}'\right] = \frac{1}{\sigma_u^2}\left[\mathbf{Q} + \psi \cdot \frac{1}{T}\boldsymbol{\iota}\boldsymbol{\iota}'\right]$$

where

$$\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$$

- Economizing on notation:

$$[W_{\tilde{x}\tilde{x}} + \psi B_{\tilde{x}\tilde{x}}] \begin{bmatrix}\hat{\mu} \\ \hat{\boldsymbol{\beta}}\end{bmatrix}_{\text{GLS}} = W_{\tilde{x}y} + \psi B_{\tilde{x}y} \tag{4.4}$$

where

$$\begin{aligned} W_{\tilde{x}\tilde{x}} &= T_{\tilde{x}\tilde{x}} - B_{\tilde{x}\tilde{x}} & W_{\tilde{x}y} &= T_{\tilde{x}y} - B_{\tilde{x}y} \\ T_{\tilde{x}\tilde{x}} &= \sum_{i=1}^{N} \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i & T_{\tilde{x}y} &= \sum_{i=1}^{N} \tilde{\mathbf{X}}_i'\mathbf{y}_i \\ B_{\tilde{x}\tilde{x}} &= \frac{1}{T}\sum_{i=1}^{N} \left(\tilde{\mathbf{X}}_i'\boldsymbol{\iota}\boldsymbol{\iota}'\tilde{\mathbf{X}}_i\right) & B_{\tilde{x}y} &= \frac{1}{T}\sum_{i=1}^{N} \left(\tilde{\mathbf{X}}_i'\boldsymbol{\iota}\boldsymbol{\iota}'\mathbf{y}_i\right) \end{aligned}$$

- Some intuition: These matrices contain the sum of squares and the sums of cross products between groups ($B_{\tilde{x}\tilde{x}}$ & $B_{\tilde{x}y}$), within groups ($W_{\tilde{x}\tilde{x}}$ & $W_{\tilde{x}y}$), and for total variation ($T_{\tilde{x}\tilde{x}}$ & $T_{\tilde{x}y}$).

- Expanding 4.4 gives

$$\begin{bmatrix} \psi NT & \psi T \sum_{i=1}^{N} \bar{\mathbf{x}}_i' \\ \psi T \sum_{i=1}^{N} \bar{\mathbf{x}}_i & \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Q} \mathbf{X}_i + \psi T \sum_{i=1}^{N} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix}_{\text{GLS}} = \begin{bmatrix} \psi NT \bar{y} \\ \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Q} \mathbf{y}_i + \psi T \sum_{i=1}^{N} \bar{\mathbf{x}}_i \bar{y}_i \end{bmatrix}$$

- Taking the partitioned inverse and solving for the parameters gives

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left[\frac{1}{T} \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Q} \mathbf{X}_i + \psi \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right]^{-1} \times \left[\frac{1}{T} \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Q} \mathbf{y}_i + \psi \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{y}_i - \bar{y}) \right]$$

$$\hat{\mu}_{\text{GLS}} = \bar{y} - \hat{\boldsymbol{\beta}}_{\text{GLS}}' \bar{\mathbf{x}}$$

- We can rewrite $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ as

$$\Delta \hat{\boldsymbol{\beta}}_{\text{BE}} + (\mathbf{I}_k - \Delta) \hat{\boldsymbol{\beta}}_{\text{CV}}$$

where

$$\Delta = \psi T \left[\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Q} \mathbf{X}_i + \psi T \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right]^{-1} \times \left[\sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right],$$

$$\hat{\boldsymbol{\beta}}_{\text{BE}} = \left[\sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right]^{-1} \left[\sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{y}_i - \bar{y}) \right].$$

- In words: the GLS estimator is a weighted average of the b/t group estimator and the w/in group estimator, w/ ψ indicating the weight given to b/t group variation. Recall

$$\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$$

  - ➢ As $\psi \to 1$, $\hat{\boldsymbol{\delta}}_{\text{GLS}} \to T_{\tilde{x}\tilde{x}}^{-1} T_{\tilde{x}y}$ (i.e., the OLS estimator). This means that little variance is explained by the unit effects.
  - ➢ As $\psi \to 0$, $\hat{\boldsymbol{\beta}}_{\text{GLS}} \to$ the w/in estimator. This happens as either
    1. the unit-specific effects dominate the disturbance u_{it}.
    2. $T \to \infty$ (intuition: the α_i are like fixed parameters since we have so much data on the T dimension).
  - ➢ GLS then is an intermediate approach b/t OLS and FE (which uses no b/t group variation).

- The variance of the GLS estimator is

$$\text{var}\left[\hat{\boldsymbol{\beta}}_{\text{GLS}}\right] = \sigma_u^2 \left[\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{Q}\mathbf{X}_i + \psi T \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'\right]^{-1}$$

- Recall the variance for the w/in group estimator:

$$\text{var}[\hat{\boldsymbol{\beta}}_{\text{CV}}] = \sigma_u^2 \left[\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{Q}\mathbf{X}_i\right]^{-1}.$$

- The difference b/t these var-cov matrices is a p.d. matrix (assuming $\psi > 0$).
- Thus, as $T \to \infty$, $\psi \to 0$, and var $\left[\sqrt{T}\hat{\boldsymbol{\beta}}_{\text{GLS}}\right] \to$ var $\left[\sqrt{T}\hat{\boldsymbol{\beta}}_{\text{CV}}\right]$ (assuming our cross-product matrices converge to finite p.d. matrices).
- Since we typically do not know σ_u^2 and σ_α^2, they must be estimated. We can do two-step GLS (i.e., obtain consistent estimates of the variance components and then plug these in to compute 2nd stage parameter estimates).
- If either $N \to \infty$ or $T \to \infty$, 2-step GLS has the same asymptotic properties as GLS w/ known variance components.

- Can use w/in and b/t group residuals to compute estimates:

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left[(y_{it} - \bar{y}_i) - \hat{\boldsymbol{\beta}}'_{\mathrm{CV}}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\right]^2}{N(T-1) - K} \tag{4.5}$$

$$\hat{\sigma}_\alpha^2 = \frac{\sum_{i=1}^{N}\left[\bar{y}_i - \tilde{\mu} - \tilde{\boldsymbol{\beta}}'\bar{\mathbf{x}}_i)\right]^2}{N - (K+1)} - \frac{1}{T}\sigma_u^2 \tag{4.6}$$

  where $\tilde{\mu}$ and $\tilde{\boldsymbol{\beta}}$ are obtained from $B_{\tilde{x}\tilde{x}}^{-1}B_{\tilde{x}y}$.

4.4 Maximum Likelihood Estimation

- RE estimates can also be computed by ML.
- To obtain the MLE, assume u_{it} and α_i are normally dist'd and start w/ the log of the likelihood function:

$$\ln L = -\frac{NT}{2}\ln 2\pi - \frac{N}{2}\ln|\mathbf{V}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i - \boldsymbol{\iota}\mu - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}_i - \boldsymbol{\iota}\mu - \mathbf{X}_i\boldsymbol{\beta})$$

- To obtain the MLE $\hat{\boldsymbol{\delta}}' = (\mu, \boldsymbol{\beta}', \sigma_u^2, \sigma_\alpha^2)$, we take partial derivatives wrt each of these parameters, set to zero and solve.
- This gives four equations that we must solve simultaneously, which can be difficult.
- Instead we can use a sequential iterative procedure, alternating back and forth b/t μ and $\boldsymbol{\beta}$ and the variance components σ_u^2 and σ_α^2.
- For N fixed and $T \to \infty$, the MLEs of $\mu, \boldsymbol{\beta}'$, and σ_u^2 are consistent and $\to$ CV estimator. The MLE of σ_α^2 is inconsistent (insufficient variation b/c of fixed N).
- With simultaneous solution of σ_α^2, it's possible to obtain a negative value. It's also possible to obtain a boundary solution, although the prob. of this $\to 0$ as either T or $N \to \infty$.

4.5 Fixed v. Random Effects

- Makes no difference for large T.
- Can make a big difference when T is fixed and N is large.
- Does it make sense to treat one source of unobserved variation as random (u_{it}) and another as fixed (α_i)?
- Conditional v. marginal inference:
  - ➢ FE is often thought of as an approach where inferences are made condit'l on effects that are in the sample.
  - ➢ RE can be thought of as making inferences that are unconditional or marginal wrt the pop. of all effects.
- Perhaps the most important consideration is whether we think the unit effects are correlated w/ explanatory variables—if so, RE is not appropriate (although may not make much difference in certain situations).

4.6 Testing between Fixed and Random Effects

- If α_i is uncorrelated with the explanatory variables $\mathbf{x}_{it}$:
  - ➢ GLS is unbiased/consistent and will achieve the Cramer-Rao lower bound (i.e., is efficient).
  - ➢ CV is unbiased/consistent but is inefficient.
- If α_i is correlated with any of the explanatory variables:
  - ➢ GLS is biased/inconsistent.
  - ➢ CV is unbiased/consistent.
- This sets us up for a Hausman test:

  H_0: $E[\alpha_i \mathbf{x}_{it}] = \mathbf{0}$; Random effects appropriate $\Rightarrow$ $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is approximately equal to $\hat{\boldsymbol{\beta}}_{\text{CV}}$ but is more efficient (has smaller standard errors).

  H_1: $E[\alpha_i \mathbf{x}_{it}] \neq \mathbf{0}$; Random effects is not appropriate $\Rightarrow$ $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ will be different from $\hat{\boldsymbol{\beta}}_{\text{CV}}$ (and inconsistent).

- In this setting, the Hausman test statistic is calculated as:

$$m = (\hat{\boldsymbol{\beta}}_{\text{CV}} - \hat{\boldsymbol{\beta}}_{\text{GLS}})' \Big(\text{var}[\hat{\boldsymbol{\beta}}_{\text{CV}}] - \text{var}[\hat{\boldsymbol{\beta}}_{\text{GLS}}] \Big)^{-1} (\hat{\boldsymbol{\beta}}_{\text{CV}} - \hat{\boldsymbol{\beta}}_{\text{GLS}})$$

- $m \sim \chi^2_K$: if m is larger than its appropriate critical value, then we reject random effects as the appropriate specification.

4.7 Application

- Let's try random effects on the Garrett data:

```
xtreg gdp oild demand corp leftlab clint, re ;

Random-effects GLS regression                   Number of obs      =       350
Group variable (i): country                     Number of groups   =        14

R-sq:  within  = 0.2225                         Obs per group: min =        25
       between = 0.0007                                        avg =      25.0
       overall = 0.1255                                        max =        25

Random effects u_i ~ Gaussian                   Wald chi2(5)       =     87.18
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         gdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        oild |  -20.44602   5.394257    -3.79   0.000    -31.01857   -9.873467
      demand |   .0075601   .0010875     6.95   0.000     .0054286    .0096915
        corp |  -1.210037    .420998    -2.87   0.004    -2.035178   -.3848961
     leftlab |  -1.256097   .4275844    -2.94   0.003    -2.094147   -.4180465
       clint |   .4653267   .1481581     3.14   0.002     .1749422    .7557112
       _cons |    5.19839   1.111886     4.68   0.000     3.019134    7.377646
-------------+----------------------------------------------------------------
     sigma_u |  .98722663
     sigma_e |  2.0405757
         rho |  .18966702   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

- Note the big differences in the coefficient values compared w/ FE (for maximum likelihood estimation, replace `re` w/ `mle`).

- Let's run a Hausman test. The commands for this model would be

```
xtreg gdp oild demand corp leftlab clint, fe ;

est store garrettfe ;

xtreg gdp oild demand corp leftlab clint, re ;

est store garrettre ;

hausman garrettfe garrettre;
```

- The results:

```
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |   garrettfe    garrettre      Difference          S.E.
-------------+----------------------------------------------------------------
        oild |   -25.59808    -20.44602       -5.152068        2.502733
     demand |    .0084949     .0075601        .0009349        .0003033
       corp |   -.2500641    -1.210037        .9599731        .5153092
    leftlab |   -1.172257    -1.256097        .0838399        .1298888
      clint |    .5030912     .4653267        .0377645        .0595242
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       15.39
                Prob>chi2 =      0.0088
```

Section 5

Non-Spherical Errors

5.1 Introduction

- Up to this point, we have assumed that our errors were spherical.
- A good deal of attention, however, has been paid to issues of non-spherical errors in panel and TSCS data (esp. the latter).

5.2 The Method of PCSEs

- Key motivation for using panel corrected standard errors (PCSEs): improve inferences made from TSCS data by taking into account the complexity of the error process, but not ask too much of data.
- Non-standard error structures (TSCS):

  1. Contemporaneous correlation: errors across cross-sect'l units are correlated due to common shocks in a given time period.

$$E(u_{it}, u_{js}) = \begin{cases} \sigma_i^2 & \text{if } i = j \text{ and } s = t \\ \sigma_{ij} & \text{if } i \neq j \text{ and } s = t \\ 0 & \text{otherwise} \end{cases}$$

  2. Panel heteroskedasticity: error var. differs across cross-sect'l units due to characteristics unique to the units.

$$E(u_{it}, u_{js}) = \begin{cases} \sigma_i^2 & \text{if } i = j \text{ and } s = t \\ 0 & \text{otherwise} \end{cases}$$

3. Serial correlation: errors w/in units are temporally correlated. E.g.,:

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

- OLS not BLUE & can produce incorrect SEs when the errors are nonspherical.
- GLS is BLUE & gives correct SEs.
- But assumes that the var-cov matrix of the errors—denoted $\text{Cov}(\mathbf{u}) = \mathbf{\Omega}$—used to weight the data is known.
- Can do FGLS using $\hat{\mathbf{\Omega}}$.
- Beck & Katz '95 *APSR* show, however, that the FGLS method advocated by Parks and Kmenta produces incorrect SEs when applied to TSCS data.
- FGLS gives overconfident SEs—does not fully take into account the variability in the estimates of the error parameters (rely too heavily on asymptotic properties).
- Beck & Katz '95 *APSR*: superior way to handle complex error structures w/ TSCS analysis is to estimate coefficients by OLS & compute PCSEs.
- Intuition: OLS with TSCS data will be unbiased but will produce incorrect standard errors.

5.3 Robust Estimation of Asymptotic Covariance Matrices

- In order to understand how PCSEs work, it's useful to go through robust estimators for cross-sectional data.
- The asymptotic covariance matrix of the OLS estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\text{var}}[\hat{\boldsymbol{\beta}}] = \frac{1}{N}\left(\frac{1}{N}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{N}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\right)\left(\frac{1}{N}\mathbf{X}'\mathbf{X}\right)^{-1} \tag{5.1}$$

  where $\boldsymbol{\Sigma} = E[\mathbf{u}\mathbf{u}'|\mathbf{X}]$.
- The problem here is how to estimate $N^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}$, since $\boldsymbol{\Sigma}$ contains $N(N+1)/2$ unknown parameters and we have only N data points.
- Fortunately, all we need is an estimator of the $K(K+1)/2$ unknown elements of

$$\text{plim } \mathbf{Q}_* = \text{plim } \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_{ij}\mathbf{x}_i\mathbf{x}_j'. \tag{5.2}$$

- $\mathbf{Q}_*$ is a matrix of the sums of squares and cross-products that involves σ_{ij} and the rows of $\mathbf{X}$.
- Since OLS $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$, then the OLS residuals $\hat{u}_i$ will be consistent for u_i, and can be used to construct estimates of σ_{ij}.
- For the case of heteroskedasticity, we want to estimate

$$\mathbf{Q}_* = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'. \tag{5.3}$$

- White (1980) has shown that for

$$\mathbf{S}_0 = \frac{1}{N}\sum_{i=1}^{N} \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i', \qquad \text{plim } \mathbf{S}_0 = \text{plim } \mathbf{Q}_* \tag{5.4}$$

  under very general conditions.

- This gives the White heteroskedasticity consistent estimator:

$$\begin{aligned} \hat{\text{var}}[\hat{\boldsymbol{\beta}}] &= \frac{1}{N}\left(\frac{1}{N}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2\mathbf{x}_i\mathbf{x}_i'\right)\left(\frac{1}{N}\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= N\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{S}_0\left(\mathbf{X}'\mathbf{X}\right)^{-1} \end{aligned} \tag{5.5}$$

- Another way to get to this result is to consider the asymptotic distribution of the OLS estimator:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{a}{\sim} N\left(\mathbf{0}, E[\mathbf{x}'\mathbf{x}]^{-1} \cdot E[u^2\mathbf{x}'\mathbf{x}] \cdot E[\mathbf{x}'\mathbf{x}]^{-1}\right) \tag{5.6}$$

  which implies

$$\text{asy. var}[\hat{\boldsymbol{\beta}}] = N^{-1}E[\mathbf{x}'\mathbf{x}]^{-1} \cdot E[u^2\mathbf{x}'\mathbf{x}] \cdot E[\mathbf{x}'\mathbf{x}]^{-1} \tag{5.7}$$

  ($\mathbf{x}$ here is the vector of explanatory variables from the population regression function and u is the population disturbance).

- To consistently estimate $E[\mathbf{x}'\mathbf{x}]$, we use the sample averages:

$$N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i = (\mathbf{X}'\mathbf{X}/N).$$

- By the law of large numbers, plim $N^{-1}\sum_{i=1}^{N} u_i^2\mathbf{x}_i'\mathbf{x}_i = E[u^2\mathbf{x}'\mathbf{x}]$. Replacing u_i with OLS residuals gives a consistent estimator of this expectation.

- Putting this altogether gives the estimator of the asymptotic variance:

$$\widehat{\text{asy. var}}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2\mathbf{x}_i'\mathbf{x}_i\right)(\mathbf{X}'\mathbf{X})^{-1} \tag{5.8}$$

- For TSCS/panel data, the basic structure is the same, but the accounting is more involved.

- One way to think about robust estimation strategies for accounting for error dependencies among cross-sectional units and time periods is to write $\text{Cov}(\mathbf{u})$ as a partitioned matrix:

$$\text{Cov}(\mathbf{u}) = \begin{bmatrix} \text{Cov}(\mathbf{u}_1) & \text{Cov}(\mathbf{u}_1,\mathbf{u}_2) & \cdots & \text{Cov}(\mathbf{u}_1,\mathbf{u}_N) \\ \text{Cov}(\mathbf{u}_1,\mathbf{u}_2) & \text{Cov}(\mathbf{u}_2) & \cdots & \text{Cov}(\mathbf{u}_2,\mathbf{u}_N) \\ \vdots & \vdots & \ddots & \vdots \\ & & & \text{Cov}(\mathbf{u}_{N-1},\mathbf{u}_N) \\ \text{Cov}(\mathbf{u}_1,\mathbf{u}_N) & \text{Cov}(\mathbf{u}_2,\mathbf{u}_N) & \cdots & \text{Cov}(\mathbf{u}_N) \end{bmatrix}.$$

  where

  - ➢ $\mathbf{u}_i$ is a $T \times 1$ vector of error terms for unit i for all time periods ($\therefore$ each block of the matrix is $T \times T$).
  - ➢ $\text{Cov}(\mathbf{u}_i)$ contains the covariances for pairs of time periods for unit i.
  - ➢ $\text{Cov}(\mathbf{u}_i,\mathbf{u}_j)$ contains the covariances for units i and j within and across time periods.

- Note: now there are potentially $NT(NT+1)/2$ unknown variance and covariance parameters to estimate in $\text{Cov}(\mathbf{u})$ with NT data points.

- Alternative robust estimators propose different ways of computing $\mathbf{X}'\widehat{\text{Cov}(\mathbf{u})}\mathbf{X}$, in part by imposing various restrictions on the elements of $\text{Cov}(\mathbf{u})$ that represent particular correlation patterns ($\widehat{\text{Cov}(\mathbf{u})}$ denotes that we have replaced the u_{it} with consistent residual estimates $\hat{u}_{it}$).

- A "cluster robust" estimator for $\text{Cov}(\boldsymbol{\beta})$ is given by substituting products of residuals (that is, $\hat{\mathbf{u}}$'s), or averages of such products in the case of constant correlations, for the non-zero products of error terms that appear in $\text{Cov}(\mathbf{u})$.

- For PCSEs, $\boldsymbol{\Omega}$ is an $NT \times NT$ band diagonal matrix with cross-sectional variances along the diagonal and contemporaneous correlations in the bands.

- For example, if $N = 2$ and $T = 3$, then

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \sigma_{12} & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 & \sigma_{12} & 0 \\ 0 & 0 & \sigma_1^2 & 0 & 0 & \sigma_{12} \\ \sigma_{12} & 0 & 0 & \sigma_2^2 & 0 & 0 \\ 0 & \sigma_{12} & 0 & 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_{12} & 0 & 0 & \sigma_2^2 \end{bmatrix}$$

- 0s in the off-diagonal elements $\Rightarrow$ no serial correlation $\Rightarrow$ need to do something else to correct for it if present.

- More generally,

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_1^2\mathbf{I}_T & \sigma_{12}\mathbf{I}_T & \cdots & \sigma_{1N}\mathbf{I}_T \\ \sigma_{21}\mathbf{I}_T & \sigma_2^2\mathbf{I}_T & \cdots & \sigma_{2N}\mathbf{I}_T \\ \vdots & & \ddots & \vdots \\ \sigma_{N1}\mathbf{I}_T & \sigma_{N2}\mathbf{I}_T & \cdots & \sigma_N^2\mathbf{I}_T \end{bmatrix} \tag{5.9}$$

- Let

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1N} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2N} \\ \vdots & & & \ddots & \vdots \\ \sigma_{1N} & \sigma_{2N} & \sigma_{3N} & \cdots & \sigma_N^2 \end{bmatrix}$$

- Use OLS residuals, denoted e_{it} for unit i at time t (in Beck and Katz's notation), to estimate the elements of $\mathbf{\Sigma}$:

$$\hat{\mathbf{\Sigma}}_{ij} = \frac{\sum_{t=1}^{T} e_{it}e_{jt}}{T}, \tag{5.10}$$

  which means the estimate of the full matrix $\hat{\mathbf{\Sigma}}$ is

$$\hat{\mathbf{\Sigma}} = \frac{\mathbf{E}'\mathbf{E}}{T}$$

  where $\mathbf{E}$ is a $T \times N$ matrix of the re-shaped $NT \times 1$ vector of OLS residuals, such that the columns contain the $T \times 1$ vectors of residuals for each cross-sectional unit (or conversely, each row contains the $N \times 1$ vector of residuals for each cross-sectional unit in a given time period) :

$$\mathbf{E} = \begin{bmatrix} e_{11} & e_{21} & \ldots & e_{N1} \\ e_{12} & e_{22} & \ldots & e_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1T} & e_{2T} & \ldots & e_{NT} \end{bmatrix}$$

Then

$$\hat{\mathbf{\Omega}} = \frac{\mathbf{E}'\mathbf{E}}{T} \otimes \mathbf{I}_T, \tag{5.11}$$

- Compute SEs using the square roots of the diagonal elements of

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \tag{5.12}$$

  where $\mathbf{X}$ denotes the $NT \times k$ matrix of stacked vectors of explanatory variables, $\mathbf{x}_{it}$.

- Intuition behind why PCSEs do well: similar to White's heteroskedasticity-consistent standard errors for cross-sect'l estimators, but better b/c take advantage of info provided by the panel structure.

- Good small sample properties confirmed by Monte Carlo studies.

- Potential issues:

  ➢ Ignores unit effects—doesn't have to, though: can just use FE residuals in eq. 5.10 or do LSDV with PCSE var-cov matrix.

  ➢ Performance of PCSEs (and any robust SE estimator) depend on whether correlation patterns in residuals are mirrored by similar patterns in the regressors.

  ➢ PCSEs solves the problems of panel heteroskedasticity and contemporaneous correlation—not serial correlation. Serial correlation must be removed before applying this fix.

  ➢ How to correct for serial correlation? Lags are recommended, but that introduces a potential problem that we'll discuss when we turn to dynamic specifications.

  ➢ Can also do Prais-Winsten.

5.4 Costs of ignoring unit effects revisited

- A dispute has arisen about the value of FE estimators when theory tells you that time-invariant variables should be included in your specification—better to just do OLS w/ PCSEs?
- We can run into trouble if unit effects are present in the data, but we ignore them for the sake of including time-invariant variables; the problem is that we don't know for sure what the true specification is.
- Methods for robust standard error estimation can help shed light on this question.
- Suppose the DGP is described by the following equation:

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i + u_{it}, \tag{5.13}$$

  where α_i indicate the unit effects and we assume u_{it} is spherical.
- If we ignore the unit effects then we are estimating the above model with the disturbance $v_{it} = \alpha_i + u_{it}$.
- If α_i is correlated with $\mathbf{x}_{it}$, then this leads to bias and inconsistency in OLS estimates of $\boldsymbol{\beta}$.
- If it is not correlated there can still be problems with inferences: relegating α_i to the disturbance term in essence induces serial correlation in the errors.
- The variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by

$$\text{var}[\hat{\boldsymbol{\beta}}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{v}\mathbf{v}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}], \tag{5.14}$$

  where $\mathbf{X}$ and $\mathbf{v}$ are the $\mathbf{x}_{it}$ and $\mathbf{v}_{it}$ stacked over all i and t.

- Let $\mathbf{\Sigma} = E[\mathbf{X}'\mathbf{v}\mathbf{v}'\mathbf{X}]$. For the case of repeated cross-sections, this can be rewritten as

$$\mathbf{\Sigma} = E\left[\sum_i \sum_j \sum_t \sum_s \mathbf{x}_{it} v_{it} v_{js} \mathbf{x}'_{js}\right] \tag{5.15}$$

- If $v_{it} = \alpha_i + u_{it}$, then

$$\mathbf{\Sigma} = E\left[\sum_i \sum_t \mathbf{x}_{it}\mathbf{x}'_{it} v_{it}^2\right] + 2E\left[\sum_i \sum_{t>s} \mathbf{x}_{it}\mathbf{x}'_{is}\alpha_i^2\right] \tag{5.16}$$

- Even if the u_{it} are spherical, the standard OLS estimator for the variance-covariance matrix will be wrong, since the second term in eq. 5.16 will be ignored if the α_i are not accounted for in the model specification.

- PCSEs will also ignore this term, leading to wrong standard errors.

- Interestingly, this presents a problem particularly for time invariant variables b/c the std. errs. for the coefficients on such variables will generally be too small, possibly leading to type I errors.

- Consider the second term of eq. 5.16 when $\mathbf{x}_{it} = (x_{it}\ z_i)$:

$$2E\left[\sum_i \sum_{t>s} \begin{bmatrix} x_{it}x_{is} & x_{it}z_i \\ x_{is}z_i & z_i z_i \end{bmatrix} \alpha_i^2\right] \tag{5.17}$$

- It is possible that the xs are uncorrelated across cross-sectional units and time periods and are uncorrelated w/ the zs $\Rightarrow$ 0s in the first diagonal element and the off-diagonals—do not contribute anything to the standard errors in expectation.

- However, the zs are perfectly correlated within $i \Rightarrow$ positive number when multiplied by $\alpha^2 \Rightarrow$ larger standard errors than what we would get from OLS.
- To confirm this analytical result, conduct MC analysis: generate the data with an explanatory variable (x_{it}) and a unit effect, but then estimate a model that replaces the unit effect with a randomly generated z_i, $z_i \perp x_{it}$; $z_i \perp \alpha_i$.

  ➢ Mimics a scenario where a researcher forsakes the FE approach to include a TINV, even though—unbeknown to the researcher—it actually has no effect.

- Estimate the model via standard OLS as well as w/ PCSEs. To verify that the problem is a kind of serial correlation induced by omission of the unit effects, we computed standard errors that are robust to serial correlation using

$$\frac{N}{N-1}\frac{NT-1}{NT-K}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i{}'\mathbf{e}_i\mathbf{e}_i'\mathbf{X}_i\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}, \qquad (5.18)$$

  where the $\mathbf{e}_i$ are the residuals from a pooled OLS regression and $\frac{N}{N-1}\frac{NT-1}{NT-K}$ is a finite sample correction.

- We could rewrite 5.18 as

$$\frac{N}{N-1}\frac{NT-1}{NT-K}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\left(\mathbf{e}\mathbf{e}'\cdot\left(\mathbf{I}_N\otimes\mathbf{J}_T\right)\right)\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}, \qquad (5.19)$$

  where $\mathbf{J}_T$ is a $T \times T$ matrix of ones and $\cdot$ indicates the Hadamard product (element-by-element multiplication).

Table 5.1: Results from Monte Carlo experiments involving time invariant variables

	No lag	One lag	Two lags
Mean OLS γ	-0.03	-0.03	-0.03
Mean Between γ	-0.02	-0.01	-0.01
Proportion of significant OLS γs	0.68	0.65	0.62
Proportion of significant OLS γs, PCSEs	0.93	0.80	0.75
Proportion of significant OLS γs, Robust	0.10	0.11	0.11
Proportion of significant Between γs	0.06	0.05	0.07
% reject H_0 of no autocorrelation	100.00	100.00	100.00

Notes: $N = 15; T = 20$. 1000 simulations for each model. The DGP does not include a time invariant variable, but the estimation equation does.

5.5 Heteroskedasticity in FE and RE models

- If we treat the α_i as random, then heteroskedasticity can occur if we have $\sigma^2_{\alpha i}$ or σ^2_{ui} or both:

$$\mathbf{V}_i = E[\mathbf{v}_i \mathbf{v}_i'] = \sigma^2_{ui} \mathbf{I}_T + \sigma^2_{\alpha i} \boldsymbol{\iota}\boldsymbol{\iota}'$$

- To do GLS, we use $\mathbf{V}_i$ instead of $\mathbf{V}$.
- Since we typically do not know either $\sigma^2_{\alpha i}$ or σ^2_{ui}, we can resort to a 2-step FGLS approach.
  - ➢ Problem: cannot get a consistent estimate for $\sigma^2_{\alpha i}$ even as $T \to \infty$; only one realization of α_i (incidental parameters).
  - ➢ Can get a consistent estimate of σ^2_{ui} as $T \to \infty$.
  - ➢ W/ finite T, cannot get consistent estimates of either $\sigma^2_{\alpha i}$ or σ^2_{ui}, even if $N \to \infty$.
  - ➢ If N **and** T are getting big, then we can get consistent estimates of σ^2_{ui}.

- To do FGLS:

  1. Run OLS or CV on $\mathbf{y}_i = \tilde{\mathbf{X}}_i\boldsymbol{\delta} + \mathbf{v}_i$ and compute the estimated residuals $\hat{\mathbf{v}}_i$.
  2. Compute $\widehat{\mathbf{V}_i}$ either by
     (a) assuming $\sigma^2_{\alpha i} = \sigma^2_\alpha \; \forall \; i$ and using

$$\hat{\sigma}^2_{ui} = \frac{1}{T-1}\sum_{t=1}^{T}\left(\hat{v}_{it} - \hat{\bar{v}}_i\right)^2, \tag{5.20}$$

     or

     (b) assuming the variance of α_i conditional on $\mathbf{x}_i$ has the same functional form across individuals $\text{var}[\alpha_i|\mathbf{x}_i] = \sigma^2\mathbf{x}_i$; $\hat{\sigma}^2_{ui}$ is estimated as in eq. 5.20 (see Roy '02 *IER* for a method of this type for heteroskedasticity of unknown form).
  3. Then do

$$\hat{\boldsymbol{\delta}}_{\text{FGLS}} = \left[\sum_{i=1}^{N}\tilde{\mathbf{X}}_i'\hat{\mathbf{V}}_i^{-1}\tilde{\mathbf{X}}_i\right]^{-1}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i\hat{\mathbf{V}}_i^{-1}\mathbf{y}_i$$

     and approximate the asymptotic var-cov matrix of $\hat{\boldsymbol{\delta}}_{\text{FGLS}}$ by

$$\left(\sum_{i=1}^{N}\tilde{\mathbf{X}}_i'\hat{\mathbf{V}}_i^{-1}\tilde{\mathbf{X}}_i\right)^{-1}$$

- Could also do a feasible weighted LS method, weighting each obs. by the reciprocal of $\hat{\sigma}_{ui}$ and then apply the CV estimator to the transformed data.

5.6 Serial Correlation in RE and FE models

- Including α_i is in essence a way to account for unobserved persistence in the data.
- Unobserved persistence can also show up in the form of serial correlation in the disturbances.
- Suppose we have the classic AR(1) disturbance:

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it} \tag{5.21}$$

  where the ε_{it} are iid w/ mean zero and variance σ^2_ε.

- If we knew ρ then we could write down a standard variance components model:

$$y_{it} - \rho y_{i,t-1} = \mu(1-\rho) + \boldsymbol{\beta}'(\mathbf{x}_{it} - \rho \mathbf{x}_{i,t-1}) + (1-\rho)\alpha_i + \varepsilon_{it} \tag{5.22}$$

  ➢ All we are doing here is substituting in $y_{i,t-1} - \mu - \boldsymbol{\beta}'\mathbf{x}_{i,t-1} - \alpha_i$ for $u_{i,t-1}$.

- Can get an asymptotically efficient estimator of $\boldsymbol{\beta}$ by doing the following:

  1. De-mean the data to take out the α_i:

  $$(y_{it} - \bar{y}_i) = \boldsymbol{\beta}'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) + (u_{it} - \bar{u}_i)$$

  2. Compute an estimate of ρ:
     (a) Use the LS residual from the de-meaned regression, or
     (b) regress $(y_{it} - \bar{y}_i)$ on $(y_{i,t-1} - \bar{y}_{i,-1})$ and $(\mathbf{x}_{i,t-1} - \bar{\mathbf{x}}_{i,-1})$ and use the coefficient on $(y_{i,t-1} - \bar{y}_{i,-1})$ as an estimate of ρ (note: $\bar{y}_{i,-1} = (1/T)\sum_{t=1}^{T} y_{i,t-1}$; we assume we have measures of y_{i0} and x_{i0}).

  3. Compute estimates of σ_ε^2 and σ_α^2:

  $$\hat{\sigma}_\varepsilon^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\{(y_{it} - \bar{y}_i) - (1-\hat{\rho})\hat{\mu} - \hat{\rho}(y_{i,t-1} - \bar{y}_{i,-1}) - \hat{\boldsymbol{\beta}}'\left[(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) - (\mathbf{x}_{i,t-1} - \bar{\mathbf{x}}_{i,-1})\hat{\rho}\right]\}^2$$

  $$\hat{\sigma}_\alpha^2 = \frac{1}{(1-\hat{\rho})^2}\cdot\frac{1}{N}\sum_{i=1}^{N}\left[\bar{y}_i - \hat{\mu}(1-\hat{\rho}) - \hat{\rho}\bar{y}_{i,-1} - \hat{\boldsymbol{\beta}}'(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{i,-1}\hat{\rho})\right]^2 - \frac{1}{T}\hat{\sigma}_\varepsilon^2$$

     Intuition: see equations for computing GLS variance estimates under assumption of spherical disturbances (i.e., eq. 4.5 and 4.6).

  4. Plug in our estimate for $\hat{\rho}$ and use $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_\alpha^2$ to compute the var-cov matrix of $(1-\hat{\rho})\alpha_i + \varepsilon_{it}$.

5.7 Robust standard error estimation with unit effects

- The properties of the FGLS estimators just described depend crucially on asymptotics—may not do well in finite samples.
- Might be better off computing some kind of robust estimator of the var-cov matrix to get our standard errors (i.e., robust to both heteroskedasticity and serial correlation).
- Robust estimators for FE models exist; although they also rely on asymptotic properties, they may do better in finite samples since they demand less of the data.

5.7.1 Arellano robust standard errors

- Let $\mathbf{y}_i^+ = \mathbf{Q}\mathbf{y}_i$, $\mathbf{X}_i^+ = \mathbf{Q}\mathbf{X}_i$, $\mathbf{u}_i^+ = \mathbf{Q}\mathbf{u}_i$.
- Arellano ('87 *Oxford Bulletin of Economics and Statistics*) suggests the following robust estimator:

$$\left(\mathbf{X}^{+\prime}\mathbf{X}^{+}\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i^{+\prime}\hat{\mathbf{u}}_i^{+}\hat{\mathbf{u}}_i^{+\prime}\mathbf{X}_i^{+}\right)\left(\mathbf{X}^{+\prime}\mathbf{X}^{+}\right)^{-1} \tag{5.23}$$

  where $\hat{\mathbf{u}}_i^+$ are the estimated residuals obtained from running OLS on the transformed equation (i.e., $\hat{\mathbf{u}}_i^+ = \mathbf{y}_i^+ - \mathbf{X}_i^+\hat{\boldsymbol{\beta}}_{\text{CV}}$).

5.7.2 Kiefer robust standard errors

- Arellano also suggests (following Kiefer) the robust estimator:

$$\left(\mathbf{X}^{+\prime}\mathbf{X}^{+}\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_{i}^{+\prime}\hat{\mathbf{\Omega}}^{+}\mathbf{X}_{i}^{+}\right)\left(\mathbf{X}^{+\prime}\mathbf{X}^{+}\right)^{-1} \tag{5.24}$$

  where

$$\hat{\mathbf{\Omega}}^{+}=\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{u}}_{i}^{+}\hat{\mathbf{u}}_{i}^{+\prime}$$

- Both of these estimators assume T fixed and $N \rightarrow \infty$.
- Monte Carlo work indicates that Arellano estimator works far better than Kiefer for TSCS data.

5.7.3 Arellano PCSEs

- Note that neither Arellano standard errors (ASEs) nor the Keifer version account for contemporaneous correlation and panel heteroskedasticity.
- Can do a hybrid of ASEs and PCSEs that account for all three types of non-sphericity.
- I.e., consider dependency in the errors that accounts for clustering within and across cross-sections, as well as across time-periods within cross-sections.
- Cov($\mathbf{u}$) is a block matrix: blocks along the diagonal contain w/in-cross-section covariances (like expression 5.23), blocks off the diagonal are diagonal matrices containing within-T covariances (like eq. (5.9) but w/o the restriction of equal covariance for all cross-section pairs.

- Following Cameron, Miller, Gelbach 2011 *J. Bus & Econ Stat*, we can construct this robust estimator as

$$\begin{aligned} &\tfrac{N}{N-1}\tfrac{NT-1}{NT-K}\left(\mathbf{X}^{+\prime}\mathbf{X}^{+}\right)^{-1}\left(\textstyle\sum_{i=1}^{N}\mathbf{X}_i^{+\prime}\hat{\mathbf{u}}_i^{+}\hat{\mathbf{u}}_i^{+\prime}\mathbf{X}_i^{+}\right)\left(\mathbf{X}^{+\prime}\mathbf{X}^{+}\right)^{-1} \\ &\quad+\tfrac{T}{T-1}\tfrac{NT-1}{NT-K}(\mathbf{X}^{+\prime}\mathbf{X}^{+})^{-1}\mathbf{X}^{+\prime}\hat{\mathbf{\Omega}}_{\text{PCSE}}\mathbf{X}^{+}(\mathbf{X}^{+\prime}\mathbf{X}^{+})^{-1} \\ &-\tfrac{NT-1}{NT-K}\left(\mathbf{X}^{+\prime}\mathbf{X}^{+}\right)^{-1}\mathbf{X}^{+\prime}\left(\hat{\mathbf{u}}^{+}\hat{\mathbf{u}}^{+\prime}\cdot\mathbf{I}_{NT}\right)\mathbf{X}^{+}\left(\mathbf{X}^{+\prime}\mathbf{X}^{+}\right)^{-1}, \end{aligned} \tag{5.25}$$

  where $^{+}$ refers to quantities subjected to the within-group transformation, $\mathbf{X}_i^{+}$ is the matrix of regressors for unit i, and $\hat{\mathbf{\Omega}}_{\text{PCSE}}$ is the same as what appears in equation 5.11, except we use residuals computed from the within-group estimator.

- This formulation demonstrates how the estimator combines the cross-section and time-period cluster robust estimators (first two lines, respectively), taking care to ensure no "double counting" of the diagonal (third line).

- MC studies show that APCSEs do very well in TSCS data of various dimensions, obviating the need to include an LDV.

5.7.4 Software

- FE robust std. errs. can be obtained in Stata by using the `robust` and `cluster` options in the `xtreg` routine.

- Douglas Miller provides an `ado` file for multi-way cluster robust standard errors (such as APCSEs) at
  `http://www.econ.ucdavis.edu/faculty/dlmiller/statfiles/`.

- In `R`: `pcse` package for PCSEs; the `plm` package implements a variety of robust SE estimators (although not clear that they include the small sample corrections); I have `R` code for ASEs and APCSEs.

5.8 Application: Garrett data

- The Garrett model of GDP growth estimated w/ OECD data is exactly the kind of case where we might expect non-spherical errors.

```
. xtpcse gdp oild demand corp leftlab clint ;

Linear regression, correlated panels corrected standard errors (PCSEs)

Group variable:   country                     Number of obs      =       350
Time variable:    year                        Number of groups   =        14
Panels:           correlated (balanced)       Obs per group: min =        25
Autocorrelation:  no autocorrelation                         avg =        25
                                                             max =        25
Estimated covariances      =       105        R-squared          =    0.1410
Estimated autocorrelations =         0        Wald chi2(5)       =     42.01
Estimated coefficients     =         6        Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |           Panel-corrected
         gdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        oild |   -15.2321   5.228694    -2.91   0.004    -25.48015   -4.984049
      demand |   .0049977   .0015394     3.25   0.001     .0019804    .0080149
        corp |  -1.139716   .2234088    -5.10   0.000    -1.577589   -.7018424
     leftlab |  -1.483549   .2755847    -5.38   0.000    -2.023685   -.9434125
       clint |   .4547183   .0839526     5.42   0.000     .2901741    .6192624
       _cons |   5.919865   .5833949    10.15   0.000     4.776432    7.063298
------------------------------------------------------------------------------
```

- PCSEs, correcting for serial correlation:

```
. xtpcse gdp oild demand corp leftlab clint,  correlation(ar1) ;

Prais-Winsten regression, correlated panels corrected standard errors (PCSEs)

Group variable:   country                     Number of obs      =       350
Time variable:    year                        Number of groups   =        14
Panels:           correlated (balanced)       Obs per group: min =        25
Autocorrelation:  common AR(1)                               avg =        25
                                                             max =        25
Estimated covariances      =       105        R-squared          =    0.1516
Estimated autocorrelations =         1        Wald chi2(5)       =     31.55
Estimated coefficients     =         6        Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |           Panel-corrected
         gdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        oild |  -13.77226   6.587739    -2.09   0.037      -26.684   -.8605317
      demand |   .0060806   .0016414     3.70   0.000     .0028635    .0092977
        corp |  -1.177445   .2934019    -4.01   0.000    -1.752502   -.6023874
     leftlab |   -1.46776   .3623476    -4.05   0.000    -2.177949    -.757572
       clint |   .4488461   .1112233     4.04   0.000     .2308525    .6668397
       _cons |   5.814019    .807692     7.20   0.000     4.230972    7.397066
-------------+----------------------------------------------------------------
         rho |   .2958842
------------------------------------------------------------------------------
```

- Fixed effects w/ robust standard errors à la Arellano:

```
. xtreg gdp oild demand corp leftlab clint, fe robust cluster(country)

Fixed-effects (within) regression               Number of obs      =       350
Group variable (i): country                    Number of groups   =        14

R-sq:  within  = 0.2315                         Obs per group: min =        25
       between = 0.0461                                        avg =      25.0
       overall = 0.0424                                        max =        25

                                                F(5,345)           =     14.29
corr(u_i, Xb)  = -0.7104                        Prob > F           =    0.0000

                                 (Std. Err. adjusted for 14 clusters in country)
------------------------------------------------------------------------------
             |               Robust
         gdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        oild |  -25.59808   14.10159    -1.82   0.093    -56.06271    4.866543
      demand |   .0084949   .0020412     4.16   0.001     .0040851    .0129047
        corp |  -.2500641   .9970869    -0.25   0.806    -2.404139    1.904011
     leftlab |  -1.172257   .3593795    -3.26   0.006    -1.948649   -.3958644
       clint |   .5030912   .1346974     3.73   0.002     .2120951    .7940873
       _cons |    1.78165   3.051424     0.58   0.569    -4.810551    8.373852
-------------+----------------------------------------------------------------
     sigma_u |  1.9296773
     sigma_e |  2.0405757
         rho |  .47208946   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

- For kicks, let's do FGLS:

```
. xtgls gdp oild demand corp leftlab clint,  panels(correlated);

Cross-sectional time-series FGLS regression

Coefficients:  generalized least squares
Panels:        heteroskedastic with cross-sectional correlation
Correlation:   no autocorrelation

Estimated covariances      =       105          Number of obs      =       350
Estimated autocorrelations =         0          Number of groups   =        14
Estimated coefficients     =         6          Time periods       =        25
                                                Wald chi2(5)       =    108.70
Log likelihood             =  -627.671          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         gdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        oild |  -14.01453   2.347816    -5.97   0.000    -18.61616   -9.412892
      demand |   .0045992   .0008527     5.39   0.000     .0029279    .0062705
        corp |  -1.107322   .1370615    -8.08   0.000    -1.375958   -.8386865
     leftlab |  -1.247823   .1348838    -9.25   0.000    -1.512191   -.9834559
       clint |   .4114256   .0479446     8.58   0.000     .3174559    .5053952
       _cons |    5.41023    .276808    19.55   0.000     4.867696    5.952763
------------------------------------------------------------------------------
```

Section 6

Dynamic Panel Models

6.1 Introduction

- How to model persistence?
  1. Lags of the dependent variable (LDV) are included as regressors
     - ➢ Account for partial adjustment of behavior over time (e.g., to reach a long-run equilibrium).
     - ➢ Account for particular factors, including exogenous shocks, that have continual effects over time (coefficient on LDV indicate whether these factors have greater impact over time or whether their impact decays and the rate at which it decays).
     - ➢ Eliminate serial correlation in the disturbance term.
     - ➢ Parsimonious way of accounting for the persistent effects of explanatory variables w/o including their lags.
  2. Individual-specific effects that do not vary over time
  3. *Dynamic panel models* employ both of these approaches: dynamics plus individual-level heterogeneity.
- Studies in economics: labor demand (Arellano and Bond '91 *Rev. of Econ. Studies*, Blundell and Bond '98 *J. of Econometrics*), firm investment (Blundell, Bond, Devereux, and Schiantarelli '92 *J. of Econometrics*), private savings rates (Loayza, Schmidt-Hebbel, and Serven '00 *Rev. of Econ. Studies*), and cigarette demand (Baltagi, Griffin, and Xiong '00 *Rev. of Econ. & Stats*).

- Studies in pol. sci.: party ID (Markus '82 *APSR*; Franklin and Jackson '83 *APSR*; Green and Palmquist '90 *AJPS*; Kabashima and Ishio '98 *Party Politics*; Green and Yoon '02 *PA*), campaign finance (Krasno, Green, and Cowden '94 *JOP*; Box-Steffensmeier and Lin '96 *PA*) and protest activity (Finkel and Mueller '98 *APSR*).
- Problems w/ pol. sci. studies:
  1. Estimate models on a period-by-period basis (i.e., they estimate a separate, cross-sectional model for each time period)—inefficient (does not take advantage of the panel structure of the data and the info it provides).
  2. Do not adequately account for individual-specific effects—one of the main motivations for doing panel analysis in the first place.
- Failure to account explicitly for unobserved individual effects $\rightarrow$ biased and inconsistent estimates.
- OLS loses the desirable properties of unbiasedness and consistency b/c at least one of the explanatory vars on the RHS will be correlated w/ disturbance.
- Even LSDV/within-group produces biased and inconsistent estimates.

6.2 Dynamic panel data and cross-sectional estimators

- OLS: both biased and inconsistent when used to estimate dynamic models with panel data.
- Consider the following representative regression model for dynamic panel data:

$$y_{i,t} = \gamma y_{i,t-1} + \beta x_{i,t} + \alpha_i + u_{i,t} \tag{6.1}$$

  where $x_{i,t}$ is an exogenous explanatory variable (scalar).
- α_i can be either fixed or random effects, since estimators have been derived for both cases.
- Assume

$$E[u_{i,t}|y_{i,t-1}, \ldots, y_{i,1}, x_{i,t}, x_{i,t-1}, \ldots, x_{i,1}] = 0. \tag{6.2}$$

- For now, also assume that the $u_{i,t}$ are serially uncorrelated and homoskedastic.

- If we have not adequately accounted for individual-specific effects, then OLS is inappropriate; eq. 6.1 becomes

$$y_{i,t} = \gamma y_{i,t-1} + \beta x_{i,t} + u^*_{i,t} \tag{6.3}$$

  where $u^*_{i,t} = \alpha_i + u_{i,t}$.

- To see why this is problematic, consider what happens if we lag eq. 6.1 one period:

$$y_{i,t-1} = \gamma y_{i,t-2} + \beta x_{i,t-1} + \alpha_i + u_{i,t-1} \tag{6.4}$$

- *By construction*, $y_{i,t-1}$ is correlated with α_i. $\therefore$, $y_{i,t-1}$ is correlated with $u^*_{i,t}$.

- For the OLS estimates of γ and β to be unbiased,

$$E[u^*_{i,t}|y_{i,t-1}, x_{i,t}] = 0, \tag{6.5}$$

- Furthermore, the performance of OLS does not improve as sample size $\uparrow$, b/c the fundamental requirement for consistency is violated.

- LSDV transformation to remove the individual effects produces biased and inconsistent estimates because correlation remains between the transformed lagged dependent variable and the transformed disturbance:

$$y_{i,t-1} - \bar{y}_{i,t-1}, \text{ where } \bar{y}_{i,t-1} = \sum_{t=2}^{T} y_{i,t-1}/(T-1) \tag{6.6}$$

$$u_{i,t} - \bar{u}_{i,t-1}, \text{ where } \bar{u}_{i,t-1} = \sum_{t=2}^{T} u_{i,t-1}/(T-1) \tag{6.7}$$

- Maybe okay as T gets big—Hurwicz/Nickell bias.

- What do do?
  1. Transform the equation to remove the individual-specific effects.
  2. Deal w/ any problems that transformation induces w/ the disturbance term (different kind of correlation between the lagged endogenous variable and the disturbance created—can use IV).

6.3 The Anderson-Hsiao Estimator

- Anderson and Hsiao ('81 *JASA*; '82 *J. Econometrics*) pointed out that first differencing eq. 6.10 eliminates the problem of correlation between the lagged endogenous variable and the individual-specific effect.
- First differencing eq. 6.1 gives

$$y_{i,t} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + \beta(x_{i,t} - x_{i,t-1}) + u_{i,t} - u_{i,t-1} \quad (6.8)$$

  which can be rewritten as

$$\Delta y_{i,t} = \gamma \Delta y_{i,t-1} + \beta \Delta x_{i,t} + \Delta u_{i,t} \quad (6.9)$$

  where Δ is the difference operator such that $\Delta y_{i,t} = y_{i,t} - y_{i,t-1}$.
- Still correlation between RHS variables and the disturbance term because $y_{i,t-1}$ in $\Delta y_{i,t-1}$ is by construction correlated with $u_{i,t-1}$ in $\Delta u_{i,t}$.
- Use IV w/ set of instruments conveniently supplied by the panel structure of the data.
- $y_{i,t-2} - y_{i,t-3}$ and $y_{i,t-2}$ are correlated with $y_{i,t-1} - y_{i,t-2}$ but not $u_{i,t} - u_{i,t-1}$ (assuming eq. 6.2 holds and there is no serial correlation).
- The same is true for $x_{i,t-2} - x_{i,t-3}$ and $x_{i,t-2}$.
- Suppose eq. 6.1 includes $y_{i,t-1}$ as the only explanatory variable:

$$y_{i,t} = \gamma y_{i,t-1} + \alpha_i + u_{i,t}. \quad (6.10)$$

- Anderson and Hsiao ('81 *JASA*) showed that

$$\gamma_{IV} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta y_{i,t}\Delta y_{i,t-2}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta y_{i,t-1}\Delta y_{i,t-2}} \tag{6.11}$$

  and

$$\gamma_{IV} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta y_{i,t} y_{i,t-2}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta y_{i,t-1} y_{i,t-2}} \tag{6.12}$$

  are consistent estimators of γ.

- Anderson-Hsiao (A-H) estimators have some problems though.

- Arellano ('89 *Econ. Letters*) shows that the estimator given in eq. 6.11 has a singularity point as well as large variances over a range of values for γ.

- Arellano and Bover ('95 *J. Econometrics*, p. 46) concluded from a Monte Carlo study that a variant of this first-difference estimator is "useless" when $N = 100$, $T = 3$ and the coefficient on the lagged endogenous variable is .8.

- Others have shown that it is inefficient b/c it neglects important information in the data.

- The subsequent improvements on A-H have built on their innovation of using IVs made available by the panel structure of the data. These studies have adopted the Generalized Method of Moments (GMM) framework to derive estimators that surmount the problems of A-H.

- GMM estimators: key intuition is that once the individual-specific effects are removed, the panel structure of the data provides a large number of IVs in the form of lagged endogenous and exogenous variables.

6.4 Review of GMM estimation

- You've done GMM before—OLS and maximum likelihood can be derived as GMM estimators—just like GLS & 2SLS.
- Main idea: from a set of basic assumptions about a DGP, we can establish population moment conditions and then use sample analogs of these moment conditions to compute parameter estimates.
- Pop. moment conditions typically involve expectations of functions of the disturbance term and explanatory variables, while the sample analogs of the population moment conditions typically take the form of sample means.
- Consider the cross-sectional regression

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i \tag{6.13}$$

where we adopt the key identifying assumption

$$E\left[\mathbf{x}_i' u_i\right] = \mathbf{0} \tag{6.14}$$

(here $\mathbf{x}_i$ is a $1 \times k$ matrix of explanatory variables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters to be estimated, and u_i is the disturbance).

- This basic assumption defines a set of moment conditions and is a weaker variant of the assumption in eq. 6.5 discussed above (violated in a dynamic panel setting).
- Substituting in for u_i, we can rewrite eq. 6.14 as

$$E\left[\mathbf{x}_i'(y_i - \mathbf{x}_i\boldsymbol{\beta})\right] = \mathbf{0} \tag{6.15}$$

to get the moment conditions in terms of observables and parameters.

- Pop. moments are estimated *consistently* with sample moments, so the next step is to write down the sample analog of eq. 6.15:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i'\left(y_i-\mathbf{x}_i\hat{\boldsymbol{\beta}}\right)=\mathbf{0} \tag{6.16}$$

  where $\hat{\boldsymbol{\beta}}$ is our estimator.

- Multiplying this out and solving for $\hat{\boldsymbol{\beta}}$ gives

$$\hat{\boldsymbol{\beta}}=\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i'y_i\right), \tag{6.17}$$

  which is identical to the equation for the OLS estimator of $\boldsymbol{\beta}$.

- We can rewrite eq. 6.17 as $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$ by stacking the $\mathbf{x}_i$ and y_i for observations $i=1,\ldots,N$ into an $N\times K$ matrix $\mathbf{X}$ and $N\times 1$ vector $\mathbf{y}$, respectively.

- What if eq. 6.14 does not hold, for example, because $x_{i,k}$ in $\mathbf{x}_i$ is correlated with u_i?

- Suppose that there are some variables $\mathbf{z}_i$ available for which

$$E\left[\mathbf{z}_i'u_i\right]=\mathbf{0} \tag{6.18}$$

  does hold and that the elements of $\mathbf{z}_i$ are partially correlated with $x_{i,k}$.

- Then $\mathbf{z}_i$ can serve as instrumental variables.

- The pop. moment conditions for the GMM estimator of $\boldsymbol{\beta}$ are

$$E\left[\mathbf{z}_i'(y_i - \mathbf{x}_i\boldsymbol{\beta})\right] = \mathbf{0} \tag{6.19}$$

which have the sample analog

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i'\left(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}\right) = \mathbf{0}. \tag{6.20}$$

- If the number of columns in $\mathbf{z}_i$ (i.e., the number of moment conditions) $>$ the number of parameters to be estimated (which is typically the case), then our equation is overidentified and there is not a closed form solution as with eq. 6.16 (which was just identified).

- To get around this problem we choose $\hat{\boldsymbol{\beta}}$ so that it minimizes the quadratic

$$\left(\sum_{i=1}^{N}\mathbf{z}_i'\left(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}\right)\right)'\mathbf{W}\left(\sum_{i=1}^{N}\mathbf{z}_i'\left(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}\right)\right), \tag{6.21}$$

where $\mathbf{W}$ is a positive semi-definite weighting matrix.

- The solution to this minimization problem does have a closed form, and with a little manipulation, we obtain

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{y}\right). \tag{6.22}$$

- Note the similarities between this GMM estimator and expressions for 2SLS estimators (note the $\mathbf{Z}$s) and GLS estimators (note the $\mathbf{W}$s).

- It can be shown that the asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is

$$\boldsymbol{\Omega} = \left(E\left[\mathbf{X}_i'\mathbf{Z}_i\right]\mathbf{W}E\left[\mathbf{Z}_i'\mathbf{X}_i\right]\right)^{-1}E\left[\mathbf{X}_i'\mathbf{Z}_i\right]\mathbf{W}\mathbf{V}\mathbf{W}E\left[\mathbf{Z}_i'\mathbf{X}_i\right]\left(E\left[\mathbf{X}_i'\mathbf{Z}_i\right]\mathbf{W}E\left[\mathbf{Z}_i'\mathbf{X}_i\right]\right)^{-1} \tag{6.23}$$

where

$$\mathbf{V} = \mathrm{Var}\left[\mathbf{Z}_i'u_i\right] = E\left[\mathbf{Z}_i'u_iu_i'\mathbf{Z}_i\right].$$

- The efficiency of the GMM estimator depends crucially on the choice of $\mathbf{W}$. In order to obtain an efficient estimator, we should choose $\mathbf{W}$ so that it makes $\mathbf{\Omega}$ as small as possible.
- The choice of $\mathbf{W}$ that does this is $\mathbf{W} = \mathbf{V}^{-1}$ (Hansen '82 *Ecta*).
- Substituting in $\mathbf{V}^{-1}$ for $\mathbf{W}$ in eq. 6.23 and canceling terms substantially simplifies the expression for the asymptotic variance, which becomes

$$\mathbf{\Omega} = (\mathbf{X}_i'\mathbf{Z}_i\mathbf{V}^{-1}\mathbf{Z}_i'\mathbf{X}_i)^{-1}. \tag{6.24}$$

- Next step: come up with a consistent estimate for $\mathbf{V}$.
- Do not get to observe u_i, so use estimated residuals produced by calculating $\hat{u}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}^*$, where $\hat{\boldsymbol{\beta}}^*$ is a first-stage, consistent estimator of $\boldsymbol{\beta}$.
- In the first stage, we typically use in eq. 6.22 the weighting matrix $\hat{\mathbf{W}}_1 = (\mathbf{Z}'\mathbf{Z})^{-1}$ to obtain $\hat{\boldsymbol{\beta}}^*$.
- The weighting matrix we use in the second stage, which is a consistent estimator for $\mathbf{V}^{-1}$, is

$$\hat{\mathbf{W}} = \hat{\mathbf{V}}^{-1} = \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_i'\hat{u}_i\hat{u}_i'\mathbf{Z}_i] \right\}^{-1}. \tag{6.25}$$

- Plugging in $\hat{\mathbf{W}}$ and $\hat{\mathbf{V}}^{-1}$ in equations (6.22) and (6.24) produces the asymptotically optimal GMM estimator.
- Note that if we assume the disturbances are homoskedastic and not serially correlated, then it would be optimal to use $(\mathbf{Z}'\mathbf{Z})^{-1}$ for $\hat{\mathbf{W}}$.

- However, using the weighting matrix given by eq. 6.25 assures that our standard errors, which we compute by taking the square root of the diagonal of

$$\hat{\mathbf{\Omega}} = (\mathbf{X}'\mathbf{Z}\hat{\mathbf{V}}^{-1}\mathbf{Z}'\mathbf{X})^{-1}, \tag{6.26}$$

  are robust to nonspherical disturbances.

- Hansen ('82 *Ecta*) shows that GMM estimators are consistent and $\overset{a}{\sim}$ normal. Thus, if an estimator can be shown to be a GMM estimator (i.e., can be derived using the GMM framework just discussed) then the "goodness" properties of consistency and asymptotic efficiency automatically follow.

  ➢ E.g, it follows that $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically distributed as $N(\boldsymbol{\beta}, \mathbf{\Omega})$.

- GMM estimators for DPD have same basic form as for cross-sectional models.

- Key feature: exploit the panel structure of the data to construct instruments that satisfy moment conditions like eq. 6.19.

6.5 A first-difference GMM estimator for dynamic panel data

- Arellano and Bond ('91 *Rev. Econ. Studies*) note that the Anderson-Hsiao estimator is inefficient because it does not use all available instruments and can be improved upon by placing it in a GMM framework—a significant contribution; led to a growth industry developing GMM estimators for DPD.

- If assume $E(u_{i,t}) = E(u_{i,t}u_{i,s}) = 0$, then the transformed residuals in eq. 6.8 have zero covariance between all $y_{i,t}$ and $x_{i,t}$ dated $t-2$ and earlier. This means we can go back through the panel from period $t-2$ to obtain appropriate instrumental variables for purging the correlation between $\Delta y_{i,t-1}$ and $\Delta u_{i,t}$. The transformed residuals satisfy a large number of moment conditions of the form

$$E[\mathbf{z}'_{i,t}\Delta u_{i,t}] = \mathbf{0}, \;\; t = 2, \ldots, T, \tag{6.27}$$

  where $\mathbf{z}_{i,t} = (x_{i,t-1}, y_{i,t-2}, x_{i,t-2}, y_{i,t-3}, x_{i,t-3}, \ldots, y_{i,1}, x_{i,1})'$ denotes the instrument set at period t.

- For notational efficiency, we can stack the time periods to write down a system of T equations for each individual:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i \tag{6.28}$$

  where

$$\mathbf{y}_i = \begin{bmatrix} \Delta y_{i,3} \\ \Delta y_{i,4} \\ \vdots \\ \Delta y_{i,T} \end{bmatrix}, \mathbf{X}_i = \begin{bmatrix} \Delta y_{i,2} & \Delta x_{i,3} \\ \Delta y_{i,3} & \Delta x_{i,4} \\ \vdots & \vdots \\ \Delta y_{i,T-1} & \Delta x_{i,T} \end{bmatrix}, \text{ and } \mathbf{u}_i = \begin{bmatrix} \Delta u_{i,3} \\ \Delta u_{i,4} \\ \vdots \\ \Delta u_{i,T} \end{bmatrix}.$$

- The set of instruments is given by the block diagonal matrix

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}_{i,3} & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_{i,4} & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{z}_{i,T} \end{bmatrix}.$$

- Note that this means that the number of instruments increases as we move through the panel. For example, if we have the simple model in eq. 6.10, then the instrument matrix becomes

$$\underset{(T-2)\times(T-2)(T-1)/2}{\mathbf{Z}_i^*} = \begin{bmatrix} y_{i,1} & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & y_{i,1} & y_{i,2} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & y_{i,1} & y_{i,2} & \cdots & y_{i,T-2} \end{bmatrix}.$$

- Hence, if $T = 5$ then we would have

$$\mathbf{Z}_i^* = \begin{bmatrix} y_{i,1} & 0 & 0 & 0 & 0 & 0 \\ 0 & y_{i,1} & y_{i,2} & 0 & 0 & 0 \\ 0 & 0 & 0 & y_{i,1} & y_{i,2} & y_{i,3} \end{bmatrix}.$$

  for our matrix of instruments.

- The vector of population moment conditions is

$$E[\mathbf{Z}_i'\mathbf{u}_i] = \mathbf{0}. \tag{6.29}$$

- The sample analog of eq. 6.29 that we use to construct an optimal GMM estimator for $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta})$ is

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{Z}_i'\mathbf{u}_i = \mathbf{0}. \tag{6.30}$$

- Let

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}, \text{ and } \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_N \end{bmatrix}$$

  (just stack the observations for all of the cross-sectional units for all periods).

- Then we can re-express eq. 6.30 as

$$\frac{1}{N}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = \mathbf{0}.$$

- The optimal GMM estimator is then given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{Z}\hat{\mathbf{V}}^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{\mathbf{V}}^{-1}\mathbf{Z}'\mathbf{y}, \tag{6.31}$$

  where $\hat{\mathbf{V}}$ is a consistent estimate of $\mathbf{V}$, the limiting variance of the sample moments $E[\mathbf{Z}_i'\mathbf{u}_i\mathbf{u}_i'\mathbf{Z}_i]$.

- If we assume conditional homoskedasticity and no autocorrelation, then the optimal choice for $\hat{\mathbf{V}}$ is $\hat{\mathbf{V}}_c = \mathbf{Z}'\mathbf{Z}$.

- But typically, want to compute 2nd stage, robust estimate. In general, the optimal choice for $\hat{\mathbf{V}}$ is

$$\hat{\mathbf{V}}_r = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i,$$

  where $\hat{\mathbf{u}}_i$ is an estimate of the vector of residuals, $u_{i,t}$, obtained from an initial consistent estimator.

- Arellano and Bond ('91 *Rev. Econ. Studies*) suggest using $\hat{\mathbf{V}}_c = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{H}\mathbf{Z}_i$ to produce the initial consistent estimator, where

$$\mathbf{H} = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \dots & 2 \end{bmatrix}.$$

- By the properties of GMM estimators, with T fixed and $N \to \infty$, $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically distributed as $N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ (Hansen '82 *Ecta*).
- The asymptotic variance $\boldsymbol{\Sigma}$ is equal to

$$\left\{E(\mathbf{X}_i'\mathbf{Z}_i)E[\mathbf{Z}_i'\mathbf{u}_i\mathbf{u}_i'\mathbf{Z}_i]^{-1}E(\mathbf{Z}_i'\mathbf{X}_i)\right\}^{-1}.$$

  A consistent estimator of the asymptotic variance is

$$\hat{\boldsymbol{\Sigma}} = \left(\mathbf{X}'\mathbf{Z}\hat{\mathbf{V}}_r^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}.$$

- Std. errs. for the first-difference estimates are obtained by taking the square root of the diagonal of $\hat{\boldsymbol{\Sigma}}$.
- If the disturbances are heteroskedastic, then the two-step estimator is more efficient. In practice, however, the asymptotic standard errors for the one-step estimator appear to be more reliable for making inferences in small samples (Arellano and Bond '91 *Rev. Econ. Studies*; Blundell and Bond '98 *J. Econometrics*).

6.6 (Possibly Superior) Alternatives to First-Differencing

- The basic first-difference estimator can perform very poorly when the autoregressive parameter $\geq .8$ (large downward biases and serious lack of precision).
- Intuition: suppose $T = 3$, then eq. 6.10 implies

$$\Delta y_{i,2} = (\gamma - 1)y_{i,1} + \alpha_i + u_{i,2}. \tag{6.32}$$

- $y_{i,1}$, the instrument for $\Delta y_{i,2}$ in the first-differenced equation is only weakly correlated with $\Delta y_{i,2}$ for values of γ close to 1, which leads to downward bias in the autoregressive parameter.
- Even if we do not expect a high degree of persistence in the data, several alternative estimators are more attractive because they can give substantial efficiency gains over the first-difference estimator.
- Can exploit additional moment conditions that arise under alternative assumptions of the DGP.
- Differences can be used as instruments to estimate equations in levels, in addition to the instruments that are available after first-differencing (Arellano and Bover '95 *J. Econometrics*; Blundell and Bond '98 *J. Econometrics*).
- Suppose we have the simple dynamic model given by eq. 6.10 and assume

$$E[\alpha_i y_{i,t}] = E[\alpha_i y_{i,s}] \tag{6.33}$$

  for all t and s. and $E[y_{i,t-1}u_{i,s}] = 0$ where $t - 1 < s$.
- Can treat this model as part of a system of two equations—one in differences and one in levels:

$$\Delta y_{i,t} = \gamma \Delta y_{i,t-1} + \Delta u_{i,t} \tag{6.34}$$

$$y_{i,t} = \gamma y_{i,t-1} + u^*_{i,t} \tag{6.35}$$

  where the second equation denotes the DGP in levels (recall that $u^*_{i,t} = \alpha_i + u_{i,t}$).

- We get moment conditions in levels in addition to the moment conditions we have with first-differencing. For example, for $T = 3$:

$$E[u^*_{i,3}\Delta y_{i,2}] = 0. \tag{6.36}$$

since

$$E[u^*_{i,3}\Delta y_{i,2}] = E[(\alpha_i + u_{i,3})(y_{i,2} - y_{i,1})] \tag{6.37}$$

$$= E[\alpha_i y_{i,2}] - E[\alpha_i y_{i,1}] + E[u_{i,3}y_{i,2}] - E[u_{i,3}y_{i,1}]. \tag{6.38}$$

- The last two expectations are zero because of our assumptions in eq. 6.2, while the difference between the first two expectations is zero because of eq. 6.33.

- The additional moment condition leads to the matrix of instruments

$$\mathbf{Z}^+_i = \begin{bmatrix} \mathbf{Z}^*_i & \mathbf{0} \\ \mathbf{0} & \Delta y_{i,2} \end{bmatrix}.$$

- MC work shows superior performance for these kinds of estimators, esp. as $\gamma \to 1$.

- Ahn and Schmidt ('95 *J. Econometrics*) point out other moment conditions that can be exploited under alternative assumptions.

- Assume that the $u_{i,t}$ are uncorrelated with each other and uncorrelated with α_i and $y_{i,0}$ (the initial value of $y_{i,t}$), then:

$$E[u_{i,T}\Delta u_{i,t}] = 0, \;\; t = 2, \ldots, T-1 \tag{6.39}$$

- Ahn and Schmidt contend that the assumptions that give rise to these additional moment conditions are not as restrictive as the assumptions that are typical with these kinds of models.

- Moment conditions hold under weaker assumptions:
  - ➢ $\text{cov}(u_{i,t}, y_{i,0})$ is the same $\forall\ t$;
  - ➢ $\text{cov}(u_{i,t}, \alpha_i)$ is the same $\forall\ t$;
  - ➢ $\text{cov}(u_{i,t}, u_{i,s})$ is the same $\forall\ t \neq s$.
- Adding the homoskedasticity assumption that $E(u_{i,t}^2)$ is the same for all T gives still more moment conditions that can be exploited.
- Using moment conditions in eq. 6.39, along with conditions such as those in eq. 6.27 which come from first-differencing, can lead to substantial gains in asymptotic efficiency.

6.6.1 Orthogonal deviations estimator

- Arellano and Bover ('95 *J. Econometrics*) propose an alternative transformation to first-differencing called forward **orthogonal deviations**: subtract the mean of all future observations in the sample for each individual.

- With this transformation, the disturbance in eq. 6.1 becomes

$$\tilde{u}_{it} = w_{it}\left(u_{i,t} - \frac{u_{i,t+1} + \ldots + u_{i,T}}{T-t}\right) \text{ for } t = 1, \ldots, T-1,$$

  where $w_{i,t} = \sqrt{(T-t)/(T-t+1)}$ is a weight that equalizes the variance of the transformed errors.

- To implement, pre-multiply the stacked levels regression (i.e., not differenced) by the forward orthogonal deviations operator

$$\mathbf{A} = diag[(T-1)/T, \ldots, 1/2]^{1/2}\mathbf{A}^{+}, \tag{6.40}$$

  where

$$\mathbf{A}^{+} = \begin{bmatrix} 1 & -(T-1)^{-1} & -(T-1)^{-1} & \ldots & -(T-1)^{-1} & -(T-1)^{-1} & -(T-1)^{-1} \\ 0 & 1 & -(T-2)^{-1} & \ldots & -(T-2)^{-1} & -(T-2)^{-1} & -(T-2)^{-1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & -1/2 & -1/2 \\ 0 & 0 & 0 & \ldots & 0 & 1 & -1 \end{bmatrix}.$$

- Intuition:

  ➢ the elements of the rows of this matrix sum to zero (i.e., will remove the individual-specific effects)

  ➢ the upper triangularity of $\mathbf{A}^{+}$ ensures the validity of the lagged endogenous variables as instruments (thus, the instrument matrix is the same as that used for the standard first difference estimator).

- More intuition: "first differences to eliminate the effects plus a GLS transformation to remove the serial correlation induced by differencing."

6.7 Finite sample considerations

- Bias/efficiency trade-off that starts to bite as T increases in size (relative to N) $\Rightarrow$ we may not want to use all available instruments.
- More instruments become available as T increases, but instruments from earlier periods in the panel become weaker the farther we progress through the panel.
- Using all of the instruments is efficient but can cause severe downward bias in GMM estimators when our sample is finite (Ziliak '97 *J. Bus. & Econ. Stats*)—overfitting.

6.8 Specification tests

- Consistency of estimators depends crucially on the assumption that the $u_{i,t}$ in eq. 6.1 are serially uncorrelated.
- If serial correlation exists, then some of our instruments will be invalid and the moment conditions used to identify parameters will not hold.
- Should test for serial correlation.

- If no serial correlation in the $u_{i,t}$ in eq. 6.1, then the first-differenced residuals should display negative 1st-order serial correlation but not 2nd-order serial correlation:
  - ➢ First differencing produces the MA(1), process $u_{i,t} - u_{i,t-1}$.
  - ➢ If our disturbances for the levels equation are $u_{i,t} - \rho u_{i,t-1}$, then differencing gives $u_{i,t} - u_{i,t-1} - \rho(u_{i,t-1} - u_{i,t-2})$
  - ➢ $y_{i,t-2}$ not valid as an instrument since it will be correlated with $u_{i,t-2}$ in the differenced disturbance term (although lagged ys at period $t-3$ and earlier remain valid instruments).
- Arellano and Bond ('91 *Rev. Econ. Studies*) give tests of 1st- and 2nd-order serial correlation based on the residuals from the two-step estimator of the first-differenced equation.
- Drawback: the complete test (i.e., for both 1st- and 2nd-order serial correlation) can be performed only for samples where $T \geq 5$, although the test statistic for 1st-order serial correlation is defined for samples where $T \geq 4$.
- Arellano and Bond omnibus test of over-identifying restrictions (or moment conditions)—helps to determine whether our assumptions about serial correlation are correct.
  - ➢ Tests whether the moment conditions over and above those needed to identify the parameters are valid.
  - ➢ Advantage: defined for samples where $T \geq 3$.
  - ➢ Disadvantages: can reject the restrictions due to forms of misspecification other than serial correlation and its asymptotic distribution is known only when the disturbance term is homoskedastic.

- Variant of the Sargan test (cf. Sargan '58 *Ecta*; Hansen '82 *Ecta*):

$$s = \hat{\mathbf{u}}'\mathbf{Z}\left(\sum_{i=1}^{N}\mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i\right)^{-1}\mathbf{Z}'\hat{\mathbf{u}} \tag{6.41}$$

  where $\hat{\mathbf{u}} = (\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_N)'$, the stacked vectors of estimated first-differenced residuals for all i and T.

- $s \overset{a}{\sim} \chi^2$ w/ df = number of columns of $\mathbf{Z}$ minus the number of explanatory variables.

- Significant χ^2 value $\Rightarrow$ overidentifying restrictions are invalid.

- Intuition: if the moment conditions given by eq. 6.29 hold, then the sample moments given by eq. 6.30 when evaluated at the parameter estimates should be close to zero, and hence the value of the quadratic function in eq. 6.41 should be small.

- Rejection of the overidentifying restrictions should lead one to reconsider the specification of the model, possibly reducing the number of instruments employed or including more lags to address serial correlation.

- Can use differences between Sargan test statistics to test the validity of additional moment conditions.

  ➢ The difference between the Sargan statistics $\sim \chi^2$ w/ df = number of new moment conditions that are used.

  ➢ A significant χ^2 value would indicate that the additional moment conditions are not valid and should not be used.

6.9 Available software

- Arellano and Bond's first-difference estimator can be implemented using the `xtabond` routine in the latest versions of Stata.
  - ➢ Can handle models such as eq. 6.1 and more complex models that include additional lags.
  - ➢ Works for data with missing observations, provided the observations are missing at random.
  - ➢ Users can select subsets of instruments in order to avoid bias from weak instruments.
  - ➢ Can also perform Sargan and serial correlation tests.
- Arellano and Bond's free DPD98 software (`GAUSS`) can do everything that `xtabond` can do, but can also compute the orthogonal deviations estimator, as well as the version of this estimator and the first-difference estimator that uses differences as instruments for the levels equation (apparently, `xtdpdsys` in `Stata 10` can do this as well).
- The `pgmm` command in the `plm` library in `R` will also estimate Arellano-Bond and has the following features:
  - ➢ One-step v. two-step estimators
  - ➢ Up to two lags of the dependent variable
  - ➢ Explicit specification of instruments
  - ➢ Computes Sargan and autocorrelation tests.

6.10 Application: Stability in Party Identification

- Debate: is party identification extremely stable, changing only at a glacial pace or can short-term forces affect the "unmoved mover"?
- Green and Palmquist ('90 *AJPS*) model party identification and short-term forces (STF) as part of the following system of equations:

$$P_{i,t} = \beta_{12} STF_{i,t} + \gamma_{11} P_{i,t-1} + u_{1i} \tag{6.42}$$

$$STF_{i,t} = \beta_{21} P_{i,t} + \gamma_{22} STF_{i,t-1} + u_{2i}. \tag{6.43}$$

- However, they estimate only the parameters in the first structural equation in the system, using $STF_{i,t-1}$ as an instrument for $STF_{i,t}$.
- Thus, they have not corrected for the problem of correlation between $P_{i,t-1}$ and individual-specific effects that would be contained in u_{1i}.
- Primary motivation: the problem of errors in measuring party identification and how such errors can lead to incorrect inferences about the effects of short-term forces on party identification.
- Wiley-Wiley method to correct for measurement error: find that the relationships between short-term forces and party identification vanish.
- Green and Yoon ('01 *PA*): the technique that Green and Palmquist use is invalid when the model assumes that intercepts vary over individuals (use Anderson-Hsiao estimator).
- Need to pool data rather than estimating separate eqn. for each t.
- Results are reported in Table 6.1.

Table 6.1: Estimates of Green and Palmquist's dynamic party identification equations

	Pooled-IV		OD			ODL		
STF variable	P^a_{t-1}	STF^a_t	P^a_{t-1}	STF^a_t	s^b	P^a_{t-1}	STF^a_t	s^b
Feeling thermometer	0.814	0.006	−0.128	−0.006	1.221	0.220	0.014	8.866
difference scores	(0.019)	(0.001)	(0.171)	(0.007)	(0.317)	(0.104)	(0.004)	(0.031)
Solve economic	0.832	0.201	−0.099	0.004	0.003	0.166	0.092	0.664
problems	(0.027)	(0.094)	(0.158)	(0.097)	(0.959)	(0.104)	(0.081)	(0.882)
Approval rating	0.858	0.106	0.026	0.032	5.746	0.203	0.148	1.079
	(0.019)	(0.037)	(0.184)	(0.087)	(0.017)	(0.101)	(0.072)	(0.782)
Carter's handling of	0.863	0.080	−0.081	0.028	3.287	0.230	0.142	4.228
inflation	(0.016)	(0.039)	(0.147)	(0.058)	(0.070)	(0.092)	(0.054)	(0.238)

[a]Standard errors in parentheses. [b] Sargan test statistic with p values in parentheses.

- Let's estimate a dynamic panel model using the Garrett data:

```
. xtabond gdp oild demand corp leftlab clint ;

Arellano-Bond dynamic panel-data estimation     Number of obs      =       322
Group variable (i): country                    Number of groups   =        14

                                                Wald chi2(6)       =    109.35

Time variable (t): year                         Obs per group: min =        23
                                                               avg =        23
                                                               max =        23

One-step results
------------------------------------------------------------------------------
       D.gdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |
          LD. |   .1842151   .0512307     3.60   0.000     .0838048    .2846254
        oild |
          D1. |  -11.75358   7.242367    -1.62   0.105    -25.94836    2.441199
      demand |
          D1. |   .0085486   .0011043     7.74   0.000     .0063843     .010713
        corp |
          D1. |  -1.389687   .8040711    -1.73   0.084    -2.965637    .1862637
     leftlab |
          D1. |  -.9370855   .5799289    -1.62   0.106    -2.073725    .1995542
       clint |
          D1. |   .2779397   .1958862     1.42   0.156    -.1059902    .6618695
       _cons |  -.0859297   .0180375    -4.76   0.000    -.1212826   -.0505768
------------------------------------------------------------------------------
Sargan test of over-identifying restrictions:
         chi2(275) =   254.61     Prob > chi2 = 0.8060

Arellano-Bond test that average autocovariance in residuals of order 1 is 0:
         H0: no autocorrelation   z =  -8.52   Pr > z = 0.0000
Arellano-Bond test that average autocovariance in residuals of order 2 is 0:
         H0: no autocorrelation   z =  -2.22   Pr > z = 0.0267
```

```
. xtabond gdp oild demand corp leftlab clint, maxldep(5);

Arellano-Bond dynamic panel-data estimation     Number of obs      =       322
Group variable (i): country                    Number of groups   =        14

                                                Wald chi2(6)       =    112.79

Time variable (t): year                         Obs per group: min =        23
                                                               avg =        23
                                                               max =        23

One-step results
------------------------------------------------------------------------------
       D.gdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |
          LD. |    .180024    .054009     3.33   0.001     .0741682    .2858797
        oild |
          D1. |  -13.05202   10.84077    -1.20   0.229    -34.29954      8.1955
      demand |
          D1. |   .0080359   .0011813     6.80   0.000     .0057207    .0103511
        corp |
          D1. |  -3.894828   1.261303    -3.09   0.002    -6.366936    -1.42272
     leftlab |
          D1. |  -1.067675   .7845493    -1.36   0.174    -2.605363    .4700135
       clint |
          D1. |   .3535734   .2631175     1.34   0.179    -.1621274    .8692741
       _cons |  -.0792513   .0199615    -3.97   0.000     -.118375   -.0401275
------------------------------------------------------------------------------
Sargan test of over-identifying restrictions:
         chi2(104) =   161.24     Prob > chi2 = 0.0003

Arellano-Bond test that average autocovariance in residuals of order 1 is 0:
         H0: no autocorrelation   z =  -8.54   Pr > z = 0.0000
Arellano-Bond test that average autocovariance in residuals of order 2 is 0:
         H0: no autocorrelation   z =  -2.00   Pr > z = 0.0458
```

```
. xtabond gdp oild demand corp leftlab clint, lags(2);

Arellano-Bond dynamic panel-data estimation     Number of obs      =       308
Group variable (i): country                    Number of groups   =        14

                                               Wald chi2(7)       =    109.75

Time variable (t): year                        Obs per group: min =        22
                                                              avg =        22
                                                              max =        22

One-step results
------------------------------------------------------------------------------
       D.gdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |
          LD. |   .1846508   .0519053     3.56   0.000     .0829182    .2863834
         L2D. |  -.0583869   .0525718    -1.11   0.267    -.1614257    .0446519
        oild |
          D1. |  -13.68715   7.373846    -1.86   0.063    -28.13962    .7653223
      demand |
          D1. |   .0082739   .0011231     7.37   0.000     .0060727    .0104751
        corp |
          D1. |  -1.366245   .8085569    -1.69   0.091    -2.950988     .218497
     leftlab |
          D1. |  -.9848365   .5880111    -1.67   0.094    -2.137317    .1676441
       clint |
          D1. |   .2950761   .2003081     1.47   0.141    -.0975207    .6876728
       _cons |  -.0947688   .0198343    -4.78   0.000    -.1336432   -.0558943
------------------------------------------------------------------------------
Sargan test of over-identifying restrictions:
         chi2(273) =   253.69     Prob > chi2 = 0.7933

Arellano-Bond test that average autocovariance in residuals of order 1 is 0:
         H0: no autocorrelation   z = -11.27   Pr > z = 0.0000
Arellano-Bond test that average autocovariance in residuals of order 2 is 0:
         H0: no autocorrelation   z =  -1.95   Pr > z = 0.0509
```

6.11 Lagged specifications for TSCS data—the method of PCSEs revisited

- A crucial assumption for the method of PCSEs: no serial correlation.
- Various ways to remove it.
  - ➢ LDV correction appears to be better than Prais-Winsten transformation (Beck & Katz '96 *PA*).
- LDV fix requires that no unit-specific effects be present in the data.
- Discussion of unit effects is largely absent from the articles that advocate and employ PCSEs.
  - ➢ $\sim$ 40% of articles that used PCSEs also used LDV.
  - ➢ Beck ('01 *Ann. Rev. of Pol. Sci.*, pp. 282–287) addresses the issue of heterogeneity in TSCS data, but does not link this discussion with the issue of serial correlation or dynamics.
    - $*$ Recommends that researchers test for unit heterogeneity in their data, however.
  - ➢ Green, Kim, & Yoon ('01 *IO*) explicitly address dangers of unmodeled unit-specific effects in TSCS data—remind us of OLS bias and inconsistency.
  - ➢ But can use FE/LSDV to fix.

- Beck & Katz ('01 *IO*): including FE for models w/ continuous dep. vars to account for unobserved heterogeneity can be worse than leaving them out.
- Bias may not be that great, in particular if the explanatory power of the unit effects is minimal.
- Loss can be great in terms of inference on substantively impt. time-invariant or slow-moving variables.
- Beck & Katz ('01 *IO*, p. 493): including lags of dep. var. can make FE less relevant (e.g., FE are similar to including a lag with a coefficient of 1).
- But even if ind. vars are not correlated with the unit-specific effects, LDVs are correlated with such effects by construction; if ind. vars are correlated with the lagged dep. var. $\rightarrow$ their coefficients can be biased and inconsistent.
- LSDV or "within group" estimator also biased & inconsistent.
- IV estimators surmount these problems, but not clear that these are appropriate for TSCS data (small N and large T).
- May work okay for studies with large N (e.g., see Blanton '99 *J. of Peace Research*; Keith '99 *J. of Peace Research*; Poe, Tate, & Keith '99 *ISQ*).

- Implications for the method of PCSEs: can see problems both with OLS point estimates and PCSEs themselves.
- If LDV is in $\mathbf{X}$ and unit effects exist but are unmodeled and thereby relegated to the disturbance term ε, then $E[\mathbf{X}'\boldsymbol{\varepsilon}]$ will not necessarily equal zero.

  ➢ LDV correlated by construction with α_i in $\boldsymbol{\varepsilon}$.
- Expectation is assumed to be zero in the proof that $\hat{\boldsymbol{\Sigma}} \overset{a.s.}{\rightarrow} \boldsymbol{\Sigma}$, and by extension $\hat{\boldsymbol{\Omega}} \overset{a.s.}{\rightarrow} \boldsymbol{\Omega}$ (see Beck & Katz '96 *PA*, pp. 32–33 and White's *Asymptotic Theory for Econometricians*. pp. 59, 165–166).
- White's proof of consistency requires that $\hat{\boldsymbol{\beta}} \overset{a.s.}{\rightarrow} \boldsymbol{\beta}$ which will not necessarily happen if $E[\mathbf{X}'\boldsymbol{\varepsilon}] \neq 0$.
- But not clear how much of a problem this is in practice.
- The theory that tells us PCSEs are reliable is grounded in their asymptotic properties, but we care mainly about small sample properties.

  ➢ But if lg. sample properties are not good, it does not bode well for the situation where we have small N or T.
- Still, including lags to eliminate serial correlation may help w/ problem of unit effects, as Beck and Katz suspect.
- Unit effects show up as serial correlation—Beck and Katz make clear researchers should test for it $\Rightarrow$ pitfall might be avoided if correct for serial correlation.
- Note that FE estimators are also biased with LDV spec, but bias goes away as T becomes large—bias $O(1/T)$.

6.11.1 Monte Carlo Studies

- Following in the B & K tradition, Kristensen, Samii, and Wawro conducted some Monte Carlo studies to check the robustness of PCSEs when both unit effects and serial correlation are present.
- Findings:
  - ➢ In a test for serial correlation, the presence of unit effects almost always leads us to reject the null of no serial correlation when no lag is included in the estimation model.
    - ∗ Do a Lagrange Multiplier test: regress estimated OLS residuals on lags of the residuals ($NT \times R^2 \sim \chi^2_1$).
    - ∗ If a lag is including, almost certain to reject null; shouldn't proceed w/o accounting for unit effects then (should tell readers this, though).
  - ➢ LDV doesn't help much in accounting for unit effects: get bias in OLS coefficients (esp. in coefficient on lag), but including a lag seems better than not including it if you do not account for unit effects.
  - ➢ As long as there is little to no correlation b/t α_i and $\mathbf{x}_{it}$ and unit effects have little explanatory power, PCSEs do fine (and better than standard FE std. errs.)
  - ➢ As $N \uparrow$ or unit effects explain more variance or correlation b/t α_i and $\mathbf{x}_{it} \uparrow$, FE w/ Arellano std. errs. (w/ a lag) outperform PCSEs (can lead to inaccurate point estimates and Type II errors).

- ➢ PCSEs do very poorly in estimating the effects of the time-invariant variable; trade-off involved in employing the FE estimator is justified when unit effects are present—OLS estimates will likely be biased (even under the most favorable conditions) to the extent of being useless for the purpose of inference.

- Jury still out on whether DPD methods will do better—some MC work that suggests that they will not do better.

6.12 Concluding thoughts

- DPD methods can be very powerful tools in solving political science puzzles.

- But data demands are great—need to think about how data will be collected if we think dynamic specification is necessary—firmly grounded in theory or an afterthought?

Section 7

Variable Coefficient Models

7.1 Introduction

- Up to this point, we've assumed that the coefficients on explanatory variables are constant across cross-sectional units.
- Theory may indicate however that slope coefficients vary across i and/or t as well.
- Fig. 7.1 indicates bias that may occur if slopes are assumed to be heterogeneous.
- A general model is:

$$y_{it} = \sum_{k=1}^{K} \beta_{kit} x_{kit} + u_{it} \tag{7.1}$$

- This model is not identified, since we have NT observations to estimate NTK parameters.

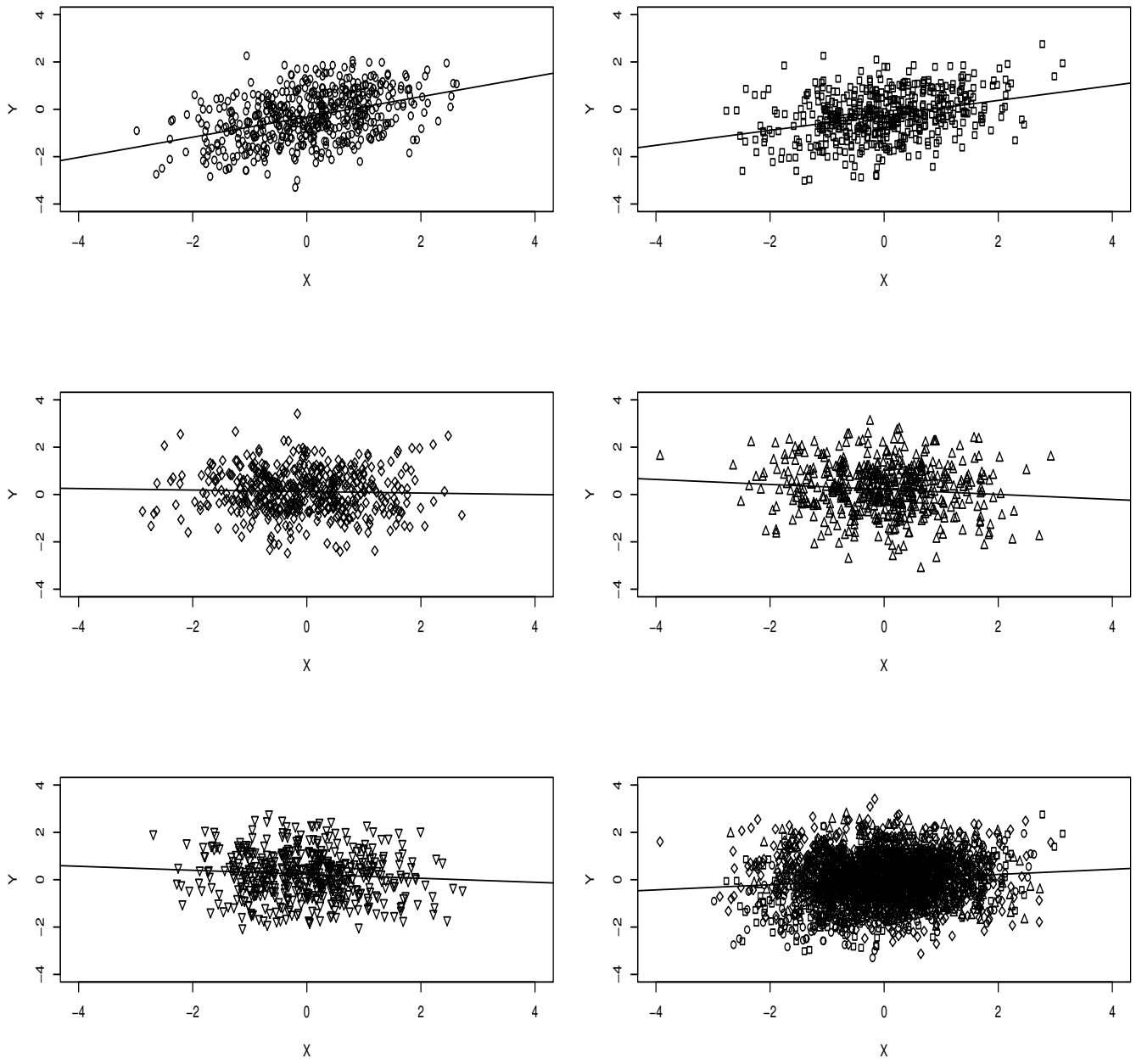


Figure 7.1: Heterogeneity Bias—variable coefficients

7.2 Cross-section specific coefficients

- Suppose we thought that the coefficients varied only across i:

$$y_{it} = \boldsymbol{\beta}_i' \mathbf{x}_{it} + u_{it} \tag{7.2}$$

- We could just run regressions on each unit separately.
- Or we could include dummy indicators for each cross-sectional unit and interact them with the explanatory variables—analogous to LSDV—still lots of parameters.
- Stack the cross-sectional regressions à la Zellner's seemingly unrelated regression model:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{X}_N \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_N \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_N \end{bmatrix} \tag{7.3}$$

- If $E[\mathbf{u}_i \mathbf{u}_j'] \neq \mathbf{0}$ then the GLS estimate of the $\boldsymbol{\beta}_i$s is more efficient than unit-by-unit OLS.
- Note that unit-by-unit OLS is an extreme approach that uses no information from other units when estimating $\boldsymbol{\beta}_i$.
- At the other extreme, we have complete pooling, where the $\boldsymbol{\beta}_i$ are constrained to be equal.

- In between these two approaches, there is "partial pooling" or "shrinkage estimators"—i.e., $\boldsymbol{\beta}_i$s vary, but are "shrunk" back toward some common mean.

  ➢ Sometimes referred to as "borrowing strength."

- We could also treat the coefficients as random variables drawn from some distribution—potentially reduces the number of parameters to be estimated:

$$\boldsymbol{\beta}_i \sim N(\boldsymbol{\beta}, \boldsymbol{\Gamma}) \tag{7.4}$$

- Can use Bayesian or classical approaches to estimation—we'll focus on the latter; for the former see Western '98 *AJPS*; not much difference w/ weak/gentle priors.

- The model in detail:

$$\begin{aligned}
y_{it} &= \boldsymbol{\beta}_i' \mathbf{x}_{it} + u_{it} \\
\boldsymbol{\beta}_i &\sim N(\boldsymbol{\beta}, \boldsymbol{\Gamma}) \\
E[(\boldsymbol{\beta}_i - \boldsymbol{\beta})|\mathbf{x}_{it}] &= \mathbf{0} \ \text{ (no systematic relationship b/t } \boldsymbol{\beta}_i \text{ and } \mathbf{x}_{it}) \\
E[u_{it}|\mathbf{x}_{it}] &= \mathbf{0} \\
E(u_{it}, u_{jt}) &= \begin{cases} \sigma_i^2 & \text{if } i = j \\ 0 & i \neq j \end{cases}
\end{aligned}$$

- If we are doing ML, then we will add $u_{it} \overset{iid}{\sim} N(0, \sigma_i^2)$

- Let $\mathbf{v}_i = \boldsymbol{\beta}_i - \boldsymbol{\beta}$ and note that $\mathbf{v}_i \sim N(\mathbf{0}, \boldsymbol{\Gamma})$.

- Rewrite the model in terms of a new, composite error term:

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + (u_{it} + \mathbf{v}_i'\mathbf{x}_{it}) = \boldsymbol{\beta}'\mathbf{x}_{it} + w_{it} \tag{7.5}$$

  Intuition: w_{it} indicates how far an individual $\boldsymbol{\beta}_i$ is from the general mean.

- Let's stack cross-sectional units in the usual fashion: $\mathbf{y}_i = \boldsymbol{\beta}'\mathbf{x}_i + \mathbf{w}_i$.

- OLS on this equation will produce consistent (although perhaps inefficient estimates):

$$\begin{aligned} E[\mathbf{w}_i|\mathbf{x}_i] &= E[(\mathbf{u}_i + \mathbf{x}_i\mathbf{v}_i)|\mathbf{x}_i] \\ &= E[\mathbf{u}_i|\mathbf{x}_i] + E[\mathbf{x}_i\mathbf{v}_i|\mathbf{x}_i] \\ &= \mathbf{0} \end{aligned}$$

- Call this pooled OLS, since we are producing only one parameter vector, $\hat{\boldsymbol{\beta}}$ (i.e., borrowing lots of strength; using all NT observations; appealing b/c of small sampling variance).

- However, this estimator is not efficient:

$$\begin{aligned} E[\mathbf{w}_i\mathbf{w}_i'] &= E[(\mathbf{u}_i + \mathbf{x}_i\mathbf{v}_i)(\mathbf{u}_i + \mathbf{x}_i\mathbf{v}_i)'] \\ &= \sigma_i^2\mathbf{I}_T + \mathbf{x}_i\boldsymbol{\Gamma}\mathbf{x}_i' \\ &= \boldsymbol{\Pi}_i \end{aligned}$$

- The var-cov matrix for the full sample is:

$$\begin{aligned} \boldsymbol{\Omega} &= \begin{bmatrix} \boldsymbol{\Pi}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Pi}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Pi}_N \end{bmatrix} \\ &= \mathbf{I}_N \otimes \boldsymbol{\Pi}_i \end{aligned}$$

- If we knew σ_i^2 and $\mathbf{\Gamma}$, could do GLS...
- Contrast w/ unit-by-unit OLS: larger sampling variance (w/ finite T) but will be consistent if there is no correlation b/t $\boldsymbol{\beta}_i$ and $\mathbf{x}_i$.

7.3 GLS/FGLS

- Assuming we have consistent estimates of σ_i^2 and $\mathbf{\Gamma}$, we could do FGLS:

$$\tilde{\boldsymbol{\beta}} = \left[\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y} \tag{7.6}$$

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

- Alternative formulation that grants more insight into the working of the GLS estimator:

$$\begin{aligned}\text{var}[\mathbf{b}_i] &= (\mathbf{x}'_i\mathbf{x}_i)^{-1}\mathbf{x}'_i\mathbf{\Pi}_i\mathbf{x}_i(\mathbf{x}'_i\mathbf{x}_i)^{-1} \\ &= \mathbf{V}_i + \mathbf{\Gamma}\end{aligned}$$

where $\mathbf{b}_i$ indicates the unit-by-unit OLS estimator and $\mathbf{V}_i = \sigma_i^2(\mathbf{x}'_i\mathbf{x}_i)^{-1}$.

- We can then rewrite the GLS estimator in eq. 7.6 as

$$\tilde{\boldsymbol{\beta}} = \sum_{i=1}^{N} \mathbf{W}_i \mathbf{b}_i \tag{7.7}$$

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

- What this says is that the GLS estimate is a weighted average of the unit-by-unit OLS estimates, where the units w/ smaller variance are given more weight

- Eq. 7.7 gives us an easy way to derive $\text{var}[\tilde{\boldsymbol{\beta}}]$ (assuming $\mathbf{b}_i$ are independent):

$$\begin{aligned} \text{var}[\tilde{\boldsymbol{\beta}}] &= \sum_{i=1}^{N} \mathbf{W}_i \text{var}[\mathbf{b}_i] \mathbf{W}_i' \\ &= \sum_{i=1}^{N} \mathbf{W}_i [\mathbf{V}_i + \boldsymbol{\Gamma}] \mathbf{W}_i' \end{aligned} \tag{7.8}$$

- Turning to GLS estimates of $\boldsymbol{\beta}_i$, we get the best linear predictor as:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_i &= [\boldsymbol{\Gamma}^{-1} + \hat{\mathbf{V}}_i^{-1}]^{-1} [\boldsymbol{\Gamma}^{-1} \tilde{\boldsymbol{\beta}} + \hat{\mathbf{V}}_i^{-1} \mathbf{b}_i] \\ &= \mathbf{A}_i \tilde{\boldsymbol{\beta}} + [\mathbf{I}_k - \mathbf{A}_i] \mathbf{b}_i \end{aligned} \tag{7.9}$$

where $\mathbf{A}_i = [\boldsymbol{\Gamma}^{-1} + \hat{\mathbf{V}}_i^{-1}]^{-1} \boldsymbol{\Gamma}^{-1}$.

- In words: our estimate is a weighted avg. b/t the pooled and unit-by-unit estimates.

- Using eq. 7.9, we can write:

$$\begin{aligned}\text{var}[\tilde{\boldsymbol{\beta}}_i] =& \mathbf{A}_i \text{var}[\tilde{\boldsymbol{\beta}}]\mathbf{A}_i' + [\mathbf{I}_k - \mathbf{A}_i]\text{var}[\mathbf{b}_i][\mathbf{I}_k - \mathbf{A}_i]' \\ &+ [\mathbf{I}_k - \mathbf{A}_i]\text{cov}[\tilde{\boldsymbol{\beta}}, \mathbf{b}_i]\mathbf{A}_i' + \mathbf{A}_i \text{cov}[\tilde{\boldsymbol{\beta}}, \mathbf{b}_i][\mathbf{I}_k - \mathbf{A}_i]'\end{aligned} \tag{7.10}$$

- In order to proceed from here, we need to figure out how to estimate $\boldsymbol{\Gamma}$ and $\mathbf{V}_i$.

- One popular alternative has been the FGLS procedure suggested by Swamy in *Statistical Inference in Random Coefficient Models*.

- Run unit-by-unit OLS in the first step to get $\mathbf{b}_i$. Then compute unit-by-unit var-cov matrices:

$$\hat{\mathbf{V}}_i = s_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$$
$$s_i^2 = \frac{\mathbf{e}_i'\mathbf{e}_i}{T-k}$$

  where $\mathbf{e}_i$ are the OLS residuals and k is the # of regressors.

- If we knew the true $\boldsymbol{\beta}_i$s we could construct the var-cov matrix for them as

$$\tilde{\boldsymbol{\Gamma}} = \frac{1}{N-1}\left(\sum_{i=1}^{N}\boldsymbol{\beta}_i\boldsymbol{\beta}_i' - N\bar{\boldsymbol{\beta}}\bar{\boldsymbol{\beta}}'\right) \tag{7.11}$$

  where $\bar{\boldsymbol{\beta}} = N^{-1}\sum_{i=1}^{N}\boldsymbol{\beta}_i$.

- $\tilde{\boldsymbol{\Gamma}} \to \boldsymbol{\Gamma}$ as N gets large. May not work well for finite T.

- Since we only have estimates of the $\boldsymbol{\beta}_i$, we need to adjust any var-cov estimate to take into account both parameter variability and sampling error (i.e., $\text{var}[\mathbf{b}_i] = \mathbf{V}_i + \boldsymbol{\Gamma}$).

- Swamy suggests using

$$\hat{\mathbf{\Gamma}} = \frac{1}{N-1}\left(\sum_{i=1}^{N}\mathbf{b}_i\mathbf{b}_i' - N\bar{\mathbf{b}}\bar{\mathbf{b}}'\right) - \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{V}}_i \tag{7.12}$$

  where $\bar{\mathbf{b}} = N^{-1}\sum_{i=1}^{N}\mathbf{b}_i$.

- Note that there is nothing that guarantees that $\hat{\mathbf{\Gamma}}$ will be positive definite; sampling variability measured by $\hat{\mathbf{V}}_i$ may swamp the measure of parameter variability estimated by the first term in 7.12.

- The standard practice is simply to drop the second term (justification: as T gets big, $\mathbf{b}_i \rightarrow \boldsymbol{\beta}_i$, and sampling variability goes away).

- For finite T, Beck and Katz suggest a "kludge" (BKK): if $\hat{\mathbf{\Gamma}}$ is negative, set it to $\mathbf{0}$ (i.e., using the fully pooled OLS estimate of $\boldsymbol{\beta}_i$).

- We could work w/ the likelihood instead, either doing classical ML or using a Bayesian approach.

- We can write the log likelihood for the varying coefficient model as

$$\begin{aligned}\ln L(\boldsymbol{\beta}_i, \sigma_i, \boldsymbol{\beta}, \mathbf{\Gamma}) = K &- \frac{T}{2}\sum_{i=1}^{N}\ln(\sigma_i^2) \\ &- \frac{1}{2}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(\mathbf{y}_i - \boldsymbol{\beta}'\mathbf{x}_i)'(\mathbf{y}_i - \boldsymbol{\beta}'\mathbf{x}_i) \\ &- \frac{N}{2}\ln|\mathbf{\Gamma}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\beta}_i - \boldsymbol{\beta})'\mathbf{\Gamma}^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta})\end{aligned} \tag{7.13}$$

  where K is some constant (i.e., contains terms not involving the parameters of interest).

- Maximizing this directly is difficult.
- "Easier" w/ Bayesian approach: specify priors for $\boldsymbol{\beta}, \boldsymbol{\Gamma}$, and σ_i^2 and then use numerical methods to calculate the full posterior (e.g., can do Markov Chain Monte Carlo à la Western '98 *AJPS*—hierarchical model).
- Beck & Katz conducted Monte Carlo experiments on these various estimators to check performance in TSCS data:
  - ➢ In terms of RMSE of β_i and its variance, FGLS does not do well w/ small T (i.e., < 40); ML and BKK both do okay; pooling also does okay.
  - ➢ Recommend against using routines in `Limdep` & `Stata` (`xtrrhh` in older versions; `xtrc` in v. 9).
- Another option: restricted ML (REML): attempt to decrease the bias affecting the maximum likelihood estimates of the variance parameters—R code available.
- Partition the likelihood into two parts—one part depending on only the variance components—free of regression coefficients; REML estimators of variance components are asymptotically the same as ML estimators.

7.4 Modeling coefficients as functions of exogenous variables

- Theory might indicate that variation in coefficients is related to measurable exogenous variables.
- If so, we can attempt to model this.
- Consider the following model:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij} \tag{7.14}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}z_j + u_{0j} \tag{7.15}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}z_j + u_{1j} \tag{7.16}$$

where

  - ➢ the ε_{ij} and $u_{.j}$ are random disturbances for which we can make a range of assumptions (iid across i and j, correlation within i, etc.).
  - ➢ For a panel/TSCS environment, we can think of i as indexing cross-sectional units (e.g., countries) and j as indexing time periods.
  - ➢ For a general multilevel environment, we could think of i as indexing micro-units (e.g., voters) and j $(= 1, \ldots, J)$ indexes macro-units (e.g., countries); hence the z_j can be thought of as contextual variables.
  - ➢ The γs are parameters to be estimated and have subscripts indicating which "level" they pertain to.

- Based on this structural model, we can write down the reduced form as

$$y_{ij} = \gamma_{00} + \gamma_{01} z_j + \gamma_{10} x_{ij} + \gamma_{11} z_j \cdot x_{ij} + u_{0j} + u_{1j} \cdot x_{ij} + \varepsilon_{ij} \quad (7.17)$$

- In substantive terms, the interaction means that the effects of micro-level variables are conditioned on the values assumed by the contextual variables.
  - ➢ E.g., country level characteristics may have different effects depending on time periods.
  - ➢ E.g., individual level characteristics may have different effects on behavioral outcomes depending on institutional configurations.
- There is a wide selection of estimators that one could use to estimate models of this kind. We'll consider a few, where the key differences across estimators relate to different assumptions about the unit level disturbances u_{0j} and u_{1j}.

7.4.1 Ordinary Least Squares

- Ignores the complicated error structure of eq. 7.17; assuming in effect that both $\text{var}[u_{0j}] = \text{var}[u_{1j}] = 0$.
- Should be less efficient than estimators that take variation from u_{0j} & u_{1j} into account; also wrong std. errs. b/c ignores heteroskedasticity of the error term.

- Can employ Huber-White consistent std. errs. (slightly modified to account for the unit level heteroskedasticity) to try to correct for this (use `cluster` option in `Stata`).

$$\frac{N-1}{N-k}\frac{J}{J-1}(X'X)^{-1}\left(\sum_{j=1}^{J}\left[\left(\sum_{i\in G_j}(e_i x_i)\right)'\left(\sum_{i\in G_j}(e_i x_i)\right)\right]\right)(X'X)^{-1} \tag{7.18}$$

  where the matrix X now includes individual (x) and unit (z) level variables.

- This var-cov matrix is consistent under arbitrary forms of heteroskedasticity within clusters as the number of J units grows large.

7.4.2 Random Effects—Restricted Maximum Likelihood

- Another option is to more fully exploit the covariance structure by assuming strict exogeneity of the disturbance terms ε_{ij}, u_{0j}, and u_{1j}:

$$\begin{aligned} L(\gamma, \Omega, \sigma^2|y_{ij}) = \int\int \frac{\sqrt{|\Omega|}}{2\pi\sigma^2}\exp(\|y_{ij} - \gamma_{00} - \gamma_{10}x_{ij} - \gamma_{01}z_j \\ - \gamma_{11}z_j \cdot x_{ij} - u_{0j} - u_{1j} \cdot x_{ij}\|^2 + u_{0j}^2 \cdot p_{11} \\ + 2u_{0j}u_{1j}p_{12} + u_{0j}^2 \cdot p_{22})du_{0j}du_{1j} \end{aligned} \tag{7.19}$$

  where

$$\Omega = \begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix}$$

  is a relative precision matrix of $u_{0\cdot}$ and $u_{1\cdot}$ compared to that of ε.

- By assuming the disturbances are independent normal variates, we can estimate Equation 7.19 by the usual (Newton) ML methods, as long as we begin with "good" starting values.

- Can use `xtmixed` in Stata; `proc mixed` is `SAS`; *lme4* package written by Bates in `R` (uses th ECME algorithm (Expectation/Conditional Maximization Either) to provide "good" starting values).

7.4.3 Random Effects—Bayesian MCMC

- This approach starts with the model fitted by the REML and then samples from the posterior distribution using Markov Chain Monte Carlo simulation.
- Priors: locally uniform for the γ parameters and inverse Wishart for the variance parameters (Ω).

7.4.4 Two-step OLS

- Could also do two-step method à la Lewis and Linzer '05 *PA*.
- Step 1: estimate a macro-unit by macro-unit regression model (i.e., estimate the model in eq. 7.14 separately for each j grouping).
  - ➢ This produces J values for the slope and intercept parameters, which then become dependent variables in the regressions given by Equations 7.15 and 7.16.
  - ➢ If the usual conditions for linear regressions are satisfied, the estimates $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ will be unbiased/consistent; but ignore any commonality across the units of analysis, so may be less efficient.

- Step 2: need to employ some method weighting each unit level observation.

  ➢ Heteroskedasticity is introduced by the fact that the dependent variable is estimated, most likely with unequal error variances across units (hence, efficiency gains from weighting).

  ➢ Saxonhouse '76 *AER* and Wooldridge '03 *AER* suggest weighting by the inverse of the std. err (downweight the observations that have more imprecise estimates).

  ➢ Hanushek '74 *Am. Statistician*, Borjas '82 *Journal of Statistical Planning and Inference* and Lewis & Linzer '05 *PA*, argue that there are multiple sources of variation implied by the model: some related to the e_{ij} term (the sampling variance), while others are related to the macro-level disturbances (e.g. u_{0j} and u_{1j}).

  ➢ Weighting by the inverse of the std. errs. neglects the existence of macro-level disturbances.

  ➢ Lewis & Linzer's MC experiments indicate that the asymptotically more efficient estimators are 10% more efficient than the OLS counterpart (w/ $J = 30$); however, std. errs. are 10% too small—tradeoff between robust inference and efficiency.

  ➢ Can use robust std. errs. instead.

7.5 Application

- In his original model, Garrett including dummies indicating certain periods where he thought different effects might have occurred:

```
. regress gdp oild demand corp leftlab clint per6673 per7479 per8084 per8690 ;

      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F(  9,   340) =   15.13
       Model |  590.718006     9   65.635334           Prob > F      =  0.0000
    Residual |  1474.47106   340  4.33667958           R-squared     =  0.2860
-------------+------------------------------           Adj R-squared =  0.2671
       Total |  2065.18906   349  5.91744717           Root MSE      =  2.0825

------------------------------------------------------------------------------
         gdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        oild |  -4.569049   4.500984    -1.02   0.311    -13.42233    4.284233
      demand |   .0028858   .0009691     2.98   0.003     .0009798    .0047919
        corp |   -.719991   .2851697    -2.52   0.012     -1.28091    -.159072
     leftlab |  -1.097082   .3564478    -3.08   0.002    -1.798203   -.3959613
       clint |   .2961244   .1152769     2.57   0.011     .0693786    .5228701
     per6673 |   1.659452   .5965629     2.78   0.006     .4860337    2.832871
     per7479 |  -.1476102   .6104749    -0.24   0.809    -1.348393    1.053173
     per8084 |  -1.028596   .6231666    -1.65   0.100    -2.254344    .1971511
     per8690 |  -.2491463   .6125551    -0.41   0.684    -1.454021    .9557287
       _cons |      4.936   .8825309     5.59   0.000     3.200092    6.671908
------------------------------------------------------------------------------
```

- Suppose we thought that the effect of different time periods extended to our slope coefficients. We could estimate a variable coefficient model w/ coefficients as functions of time period indicators.
- Let's keep this simple and make our coefficients functions of the period indicator for 1966–1973.

- Do something like this in `Stata`:

```
. xtmixed gdp oild demand corp leftlab clint || per6673: oild demand
> corp leftlab clint, cov(id) ;

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log restricted-likelihood = -778.01546
Iteration 1:   log restricted-likelihood = -777.82157  (backed up)
Iteration 2:   log restricted-likelihood = -777.81806
Iteration 3:   log restricted-likelihood = -777.81805

Computing standard errors:

Mixed-effects REML regression                   Number of obs      =       350
Group variable: per6673                         Number of groups   =         2

                                                Obs per group: min =       112
                                                               avg =     175.0
                                                               max =       238


                                                Wald chi2(5)       =     33.75
Log restricted-likelihood = -777.81805          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         gdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        oild |  -10.68686   4.495998    -2.38   0.017    -19.49885   -1.874866
      demand |   .0053231   .0030293     1.76   0.079    -.0006142    .0112603
        corp |  -1.033495   .2941646    -3.51   0.000    -1.610047    -.456943
     leftlab |  -1.456091    .370648    -3.93   0.000    -2.182547   -.7296338
       clint |   .4204303    .119154     3.53   0.000     .1868927    .6539679
       _cons |   5.754773   .7098348     8.11   0.000     4.363522    7.146023
------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
per6673: Identity            |
          sd(oild.._cons)(1) |   .0040601   .0032621      .0008407    .0196086
-----------------------------+------------------------------------------------
                sd(Residual) |   2.188893   .0836016       2.03102    2.359038
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =    21.88 Prob >= chibar2 = 0.0000
(1) oild demand corp leftlab clint _cons
```

- In SAS, can do something like:

```
proc mixed data=sasdata.garrum6;
  model gdp = oild demand corp leftlab clint per6673*oild
  per6673*demand per6673*corp per6673*leftlab per6673*clint /  solution;
  random intercept per6673 / type=vc subject=country ;
run;
```

- Produces:

The Mixed Procedure

Model Information

Data Set	SASDATA.GARRUM6
Dependent Variable	GDP
Covariance Structure	Variance Components
Subject Effect	COUNTRY
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Dimensions

Covariance Parameters	3
Columns in X	11
Columns in Z Per Subject	2

Subjects	14
Max Obs Per Subject	25

Number of Observations

Number of Observations Read	350
Number of Observations Used	350
Number of Observations Not Used	0

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	1515.67161589	
1	3	1472.04907393	0.00026453
2	1	1471.92160942	0.00001562
3	1	1471.91470245	0.00000007
4	1	1471.91467342	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	COUNTRY	0.6358
PER6673	COUNTRY	1.2065
Residual		3.4642

Fit Statistics

-2 Res Log Likelihood	1471.9
AIC (smaller is better)	1477.9
AICC (smaller is better)	1478.0
BIC (smaller is better)	1479.8

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	4.8078	1.0650	13	4.51	0.0006
OILD	-4.4086	5.5702	312	-0.79	0.4293
DEMAND	0.006484	0.001091	312	5.94	<.0001
CORP	-1.2549	0.4079	312	-3.08	0.0023
LEFTLAB	-1.2267	0.4459	312	-2.75	0.0063
CLINT	0.4442	0.1492	312	2.98	0.0031
OILD*PER6673	65.3800	126.53	312	0.52	0.6057
DEMAND*PER6673	-0.00398	0.002960	312	-1.35	0.1794
CORP*PER6673	1.1413	0.3692	312	3.09	0.0022
LEFTLAB*PER6673	0.5249	0.7015	312	0.75	0.4548
CLINT*PER6673	-0.3050	0.1936	312	-1.58	0.1161

Section 8

Models for Qualitative Dependent Variables

8.1 Introduction

- So far, we have considered models where the dependent variable is continuous.
- Methods have been developed for models where the dependent variable is dichotomous, polychotomous, censored, truncated, etc.
- Previous concerns (e.g., fixed or random effects, T v. N) come into play with non-continuous dependent variables; there are also some new challenges.

8.2 Dichotomous Dependent Variables

- The basic set up of the repeated observations model for dichotomous dependent variables is similar to the standard models.

$$y_{it}^* = \beta \mathbf{x}_{it} + \alpha_i + u_{it}, \tag{8.1}$$

where

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^* > 0 \\ 0 & \text{if } y_{it}^* \leq 0 \end{cases}$$

and $i = 1, \ldots, N$ and $t = 1, \ldots, T$, and u_{it} is assumed to be iid with mean zero and variance σ_u^2.

- The choice we make about the distribution for the disturbance term matters a lot w/ dichotomous dep. vars; still based on our beliefs about the correlation between the explanatory variables and the individual specific effect, but also has implications for estimation approach.
- If $\alpha_i \perp \mathbf{x}_{it}$, estimate a random effects model, assuming $\alpha \sim IID(0, \sigma_\alpha^2)$.
- If correlation between α_i and $\mathbf{x}_{it}$, estimate a fixed effects model (again, α_i are fixed parameters to be estimated).
- If $T \to \infty$, then it is possible to get consistent estimates of $\boldsymbol{\beta}$ and α_i.
- However, if T is fixed and $N \to \infty$, then we have the incidental parameters problem—i.e., since the number of parameters increases with N, we cannot consistently estimate α_i for fixed T.
- We can show mathematically that the inconsistency in α_i is transmitted to $\boldsymbol{\beta}$; but may not be that bad in practice (only with really large N).
- The transformations that we perform in the linear regression case (viz., subtracting off within-group times means of the variables, differencing) are not valid with a qualitative limited dependent variable model b/c of nonlinearity of such models.

8.3 Fixed Effect Logit

- Chamberlain ('80 *Rev. of Econ. Studies*) has derived a conditional maximum likelihood estimator (CMLE) that works in a way similar to the w/in transformation. The conditional likelihood function can be written as

$$L = \prod_{i=1}^{N} \Pr\left(y_{i1}, \ldots, y_{iT} | \sum_{t=1}^{T} y_{it} \right)$$

- Consider the case where $T = 2$. The unconditional likelihood is

$$L = \prod_{i=1}^{N} \Pr(y_{i1}) \Pr(y_{i2})$$

- Note that:

$$\Pr[y_{i1} = 0, y_{i2} = 0 | y_{i1} + y_{i2} = 0] = 1$$
$$\Pr[y_{i1} = 1, y_{i2} = 1 | y_{i1} + y_{i2} = 2] = 1$$

  which means these probabilities add no information to the conditional log likelihood so we can ignore them.

- But

$$\Pr[y_{i1}=0, y_{i2}=1 | y_{i1}+y_{i2}=1] = \frac{\Pr[y_{i1}=0, y_{i2}=1 \text{ and } y_{i1}+y_{i2}=1]}{\Pr[y_{i1}+y_{i2}=1]}$$
$$= \frac{\Pr[y_{i1}=0, y_{i2}=1 \text{ and } y_{i1}+y_{i2}=1]}{\Pr[y_{i1}=0, y_{i2}=1]+\Pr[y_{i1}=1, y_{i2}=0]}$$

- If we assume that the data follow a logistic distribution then we can rewrite this as

$$\frac{\frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}\frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}}{\frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}\frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})} + \frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}\frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}}$$

  which simplifies to

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{i2})}{\exp(\boldsymbol{\beta}'\mathbf{x}_{i1})+\exp(\boldsymbol{\beta}'\mathbf{x}_{i2})}$$

- The expression for the remaining probability is similarly derived. These probabilities then constitute the conditional log-likelihood function, which can be maximized using standard techniques.

- This can be extended to T of arbitrary size but the computations are excessive for $T > 10$.

- Can't use standard specification test like LR for checking unit heterogeneity b/c likelihoods are not comparable (CML uses a restricted data set).
- But can use this estimator to do a Hausman test for the presence of individual effects.
- Intuition: in the absence of individual specific effects, both the Chamberlain estimator and the standard logit maximum likelihood estimator are consistent, but the former is inefficient. If individual specific effects exist, then the Chamberlain estimator is consistent while the standard logit MLE is inconsistent.

  ➢ Inefficiency is due to loss of information/throwing away observations.

- We compute the following statistic for the test:

$$\chi^2_k = \left(\hat{\boldsymbol{\beta}}_{\text{CML}} - \hat{\boldsymbol{\beta}}_{\text{ML}}\right)' \left[\mathbf{V}_{\text{CML}} - \mathbf{V}_{\text{ML}}\right]^{-1} \left(\hat{\boldsymbol{\beta}}_{\text{CML}} - \hat{\boldsymbol{\beta}}_{\text{ML}}\right)$$

  If we get a significant χ^2 value we reject the null of no individual specific effects.

- If $\mathbf{V}_{\text{ML}} > \mathbf{V}_{\text{CML}}$, assume zero χ^2 statistic.

8.4 Application: Unionization of Women in the U.S. (from Stata manual)

- Chamberlain's fixed effects logit model can be estimated using `xtlogit` with the `fe` option.
- Sample is a panel (unbalanced) of 4,434 women from 1970–1988.
- Dep. var. is a dummy indicating union membership.
- Ind. vars include: age, years of schooling, amount of time spent living outside an SMSA, years spent living in the South, and an interaction between this variable and year.
- First, let's estimate a pooled logit model.

```
. logit union age grade not_smsa south southXt;

Iteration 0:   log likelihood =  -13864.23
Iteration 1:   log likelihood = -13550.511
Iteration 2:   log likelihood =  -13545.74
Iteration 3:   log likelihood = -13545.736

Logit estimates                                   Number of obs   =      26200
                                                  LR chi2(5)      =     636.99
                                                  Prob > chi2     =     0.0000
Log likelihood = -13545.736                       Pseudo R2       =     0.0230

------------------------------------------------------------------------------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0099931   .0026737     3.74   0.000     .0047527    .0152335
       grade |   .0483487   .0064259     7.52   0.000     .0357541    .0609432
    not_smsa |  -.2214908   .0355831    -6.22   0.000    -.2912324   -.1517493
       south |  -.7144461   .0612145   -11.67   0.000    -.8344244   -.5944678
     southXt |   .0068356   .0052258     1.31   0.191    -.0034067    .0170779
       _cons |  -1.888256    .113141   -16.69   0.000    -2.110009   -1.666504
```

- Save the results for later use:

```
. est store unionpool
```

- Now let's estimate the fixed effects model, accounting for unit heterogeneity (and save the results).

```
. xtlogit union age grade not_smsa south southXt, i(id) fe;

note: multiple positive outcomes within groups encountered.
note: 2744 groups (14165 obs) dropped due to all positive or
      all negative outcomes.
Iteration 0:   log likelihood = -4541.9044
Iteration 1:   log likelihood = -4511.1353
Iteration 2:   log likelihood = -4511.1042

Conditional fixed-effects logit                 Number of obs      =     12035
Group variable (i) : idcode                     Number of groups   =      1690

                                                Obs per group: min =         2
                                                               avg =       7.1
                                                               max =        12

                                                LR chi2(5)         =     78.16
Log likelihood  = -4511.1042                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0079706   .0050283     1.59   0.113    -.0018848    .0178259
       grade |   .0811808   .0419137     1.94   0.053    -.0009686    .1633302
    not_smsa |   .0210368    .113154     0.19   0.853    -.2007411    .2428146
       south |  -1.007318   .1500491    -6.71   0.000    -1.301409   -.7132271
     southXt |   .0263495   .0083244     3.17   0.002      .010034    .0426649
------------------------------------------------------------------------------


. est store unionfe
```

- To run the Hausman test, do

```
. hausman unionfe unionpool
```

- This gives:

```
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)   sqrt(diag(V_b-V_B))
             |     Prior       Current       Difference        S.E.
-------------+------------------------------------------------------------
         age |   .0079706     .0099931        -.0020225     .0042586
       grade |   .0811808     .0483487         .0328321     .0414182
    not_smsa |   .0210368    -.2214908         .2425276     .1074136
       south |  -1.007318    -.7144461        -.2928718     .1369945
     southXt |   .0263495     .0068356         .0195139     .0064797
------------------------------------------------------------------------------
                 b = less efficient estimates obtained previously from clogit
                 B = fully efficient estimates obtained from logit

    Test:  Ho:  difference in coefficients not systematic

                 chi2(  5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                           =    20.50
                           Prob>chi2 =     0.0010
```

8.5 Random Effects Probit

- The probit model does not lend itself to the fixed effects treatment because there is no way to sweep out the individual specific effects. But we can estimate a probit model if we assume random effects.
- Let

$$\varepsilon_{it} = \alpha_i + u_{it}$$

and assume $\alpha_i \sim N(0, \sigma_\alpha^2)$, $u_{it} \sim N(0, \sigma_u^2)$, and α_i and u_{it} are independent of each other. Then

$$\text{var}[\varepsilon_{it}] = \sigma_u^2 + \sigma_\alpha^2 = 1 + \sigma_\alpha^2$$

and

$$\text{corr}[\varepsilon_{it}, \varepsilon_{is}] = \rho = \frac{\sigma_\alpha^2}{1 + \sigma_\alpha^2}$$

for $t \neq s$. This implies $\sigma_\alpha^2 = \rho/(1-\rho)$.

- We can write the probability associated with an observation as

$$\Pr[y_{it}] = \int_{-\infty}^{q_{it}\boldsymbol{\beta}'\mathbf{x}_{it}} f(\varepsilon_{it}) d\varepsilon_{it} = \Phi[q_{it}\boldsymbol{\beta}'\mathbf{x}_{it}]$$

where $q_{it} = 2y_{it} - 1$.

- Because of the α_i, the T observations for i are jointly normally distributed. The individual's contribution to the likelihood is

$$\begin{aligned} L_i &= \Pr[y_{i1}, y_{i2}, \ldots, y_{iT}] \\ &= \int_{-\infty}^{q_{i1}\boldsymbol{\beta}'\mathbf{x}_{i1}} \int_{-\infty}^{q_{i2}\boldsymbol{\beta}'\mathbf{x}_{i2}} \cdots \int_{-\infty}^{q_{iT}\boldsymbol{\beta}'\mathbf{x}_{iT}} f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT}) d\varepsilon_{iT} \cdots d\varepsilon_{i2} d\varepsilon_{i1} \end{aligned}$$

- Rather than evaluating multiple integrals, a simplification is possible. Consider the joint density:

$$f(\varepsilon_{i1}, \ldots, \varepsilon_{iT}, \alpha_i) = f(\varepsilon_{i1}, \ldots, \varepsilon_{iT}|\alpha_i) f(\alpha_i)$$

- We can then integrate over α_i:

$$f(\varepsilon_{i1}, \ldots, \varepsilon_{it}) = \int_{-\infty}^{\infty} f(\varepsilon_{i1}, \ldots, \varepsilon_{iT}|\alpha_i) f(\alpha_i) d\alpha_i$$

- Conditioned on α_i, the ε_is are independent:

$$f(\varepsilon_{i1}, \ldots, \varepsilon_{it}) = \int_{-\infty}^{\infty} \prod_{t=1}^{T} f(\varepsilon_{it}|\alpha_i) f(\alpha_i) d\alpha_i \tag{8.2}$$

- Treat eq. 8.2 as function of α_i:

$$L_i = \int_{-\infty}^{\infty} \frac{1}{\sigma_\alpha \sqrt{2\pi}} e^{-\frac{\alpha_i^2}{2\sigma_\alpha^2}} g(\alpha_i) d\alpha_i \tag{8.3}$$

- Let $r_i = \frac{\alpha_i}{\sigma_\alpha \sqrt{2}}$, which implies $\alpha_i = \sigma_\alpha \sqrt{2} r_i = \theta r_i$ and $d\alpha_i = \theta dr_i$.
- Making the change of variable gives

$$L_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-r_i^2} g(\theta r_i) dr_i \tag{8.4}$$

- Working back to the probit model, we get i's contribution to the likelihood as

$$L_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-r_i^2} \left\{ \prod_{t=1}^{T} \Phi[q_{it}(\boldsymbol{\beta}' \mathbf{x}_{it} + \theta r_i)] \right\} dr_i \tag{8.5}$$

- Note that $\theta = \sqrt{\frac{2\rho}{1-\rho}}$.

- Things to note:
  - ➢ Still assuming $\alpha_i \perp \mathbf{x}_{it}$. We are also assuming that the within-cross section correlation is the same across all time periods.
  - ➢ ρ can be interpreted as the proportion of the variance contributed by the unit effects.
  - ➢ We can test for unit heterogeneity by checking the statistical significance of ρ. One way to do this is with a likelihood ratio ratio test of the random effects probit and pooled probit models.
  - ➢ The standard way to evaluate the integral in the likelihood is by Gauss-Hermite quadrature. This raises some concerns about how the size of T and N affect the accuracy of the quadrature approximation, and some checks of the performance of the approximation are in order.
  - ➢ Stata's `xtprobit` command can be used to estimate this model.
  - ➢ We could derive this model for the logistic distribution rather than the normal distribution.

8.6 Application: RE Probit for PAC contributions and roll call votes

- A major concern is the effects of campaign contributions on the behavior of members of the U.S. Congress.
- Assessing these effects is complicated because it is methodologically difficult to account for members' predispositions to vote in favor of PACs' interests.
- Estimating a RE probit model can help overcome this problem because it enables us to account for individual specific effects, such as the predisposition to vote for or against a particular piece of legislation, which are too costly or impossible to measure.
- The data set includes PAC contributions and roll-call votes that are of particular interest to certain types of PACs (see Wawro '01 $AJPS$).
- Let's estimate a pooled model and RE probit for USCC votes in the 102nd Congress, 2nd Session.

The pooled model:

```
. probit vote labor corp unemp;

Iteration 0:   log likelihood =  -1385.746
Iteration 1:   log likelihood = -1215.6587
Iteration 2:   log likelihood = -1214.1783
Iteration 3:   log likelihood =  -1214.178

Probit estimates                                  Number of obs   =       2002
                                                  LR chi2(3)      =     343.14
                                                  Prob > chi2     =     0.0000
Log likelihood =  -1214.178                       Pseudo R2       =     0.1238

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       labor |  -.1197816   .0075367   -15.89   0.000    -.1345532     -.10501
        corp |   .0970168    .009864     9.84   0.000     .0776837    .1163499
       unemp |  -6.325817   1.252524    -5.05   0.000    -8.780719   -3.870916
       _cons |   .3775498   .1226202     3.08   0.002     .1372186     .617881
------------------------------------------------------------------------------
```

From these results, we would infer a strong relationship between contributions and votes.

And now the RE probit model:

```
. xtprobit vote labor corp unemp, i(icpsr) nolog;

Random-effects probit                           Number of obs      =      2002
Group variable (i) : icpsr                      Number of groups   =       286

Random effects u_i ~ Gaussian                   Obs per group: min =         7
                                                               avg =       7.0
                                                               max =         7

                                                Wald chi2(3)       =     44.06
Log likelihood  = -1158.7298                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       labor |  -.0606706   .0124172    -4.89   0.000    -.0850078   -.0363333
        corp |   .0494033   .0140651     3.51   0.000     .0218362    .0769705
       unemp |  -8.208123   2.193177    -3.74   0.000    -12.50667   -3.909576
       _cons |   .6124466   .2044377     3.00   0.003     .2117559    1.013137
-------------+----------------------------------------------------------------
    /lnsig2u |  -.5200013   .1903423                     -.8930653   -.1469373
-------------+----------------------------------------------------------------
     sigma_u |   .7710511   .0733818                      .6398429    .9291653
         rho |   .3728519   .0445084                      .2904777    .4633316
------------------------------------------------------------------------------
Likelihood ratio test of rho=0: chibar2(01) =   110.90 Prob >= chibar2 = 0.000
```

These results indicate a less strong—but still statistically significant—relationship between contributions and votes.

8.7 Correlated Random Effects Probit

- The big drawback of the random effects model is the unattractive assumption that the α_i and $\mathbf{x}_i$ are uncorrelated.
- Chamberlain has proposed a correlated random effects (CRE) model that gets around this problem by assuming a specific functional relationship between α_i and $\mathbf{x}_i$. That is,

$$\alpha_i = \sum_{t=1}^{T} \mathbf{a}_t' \mathbf{x}_{it} + \eta_i = \mathbf{a}' \mathbf{x}_i + \eta_i$$

  where $\mathbf{a}' = (\mathbf{a}_1', \ldots, \mathbf{a}_T')$, $\mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, and η_i is normally distributed (with mean zero and variance σ_η^2) and is independent of the $\mathbf{x}_{it}$.
- The $\mathbf{a}_t$ are vectors of parameters to be estimated and capture the nature of the relationship between α_i and $\mathbf{x}_{it}$.
- Equation (8.1) then becomes

$$y_{it}^* = \boldsymbol{\beta}' \mathbf{x}_{it} + \sum_{t=1}^{T} \mathbf{a}_t' \mathbf{x}_{it} + \varepsilon_{it}$$

  where $\varepsilon_{it} = \eta_i + u_{it}$.

- Estimation of the CRE model proceeds in a sequential manner:

  1. Estimate separate probit equations by maximum likelihood for each time period, regressing the dependent variable in each period on all of the leads and lags of the explanatory variables.
  2. Stack the estimates from each of these probits into a vector $\hat{\pi}$ and construct the joint covariance matrix of all of the estimates. The vector $\hat{\pi}$ is a vector of reduced form estimates.
  3. Use a minimum distance estimator to impose restrictions on $\hat{\pi}$ to back out estimates of the structural parameters $\boldsymbol{\beta}$ and $\mathbf{a}_t$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{a}')$ and choose $\boldsymbol{\theta}$ to minimize

  $$[\hat{\pi} - \mathbf{f}(\boldsymbol{\theta})]' \hat{\boldsymbol{\Omega}}^{-1} [\hat{\pi} - \mathbf{f}(\boldsymbol{\theta})],$$

  where $\hat{\boldsymbol{\Omega}}$ is an estimate of the asymptotic variance-covariance matrix for the reduced-form estimates.
  4. We then conduct our standard hypothesis tests to make inferences about the effects of the variables of interest.
  5. Marginal effects can be determined by simulating probabilities:

$$\Pr(y_{it} = 1) = \Phi \left[(1 + \sigma_\eta^2)^{-1/2} (\boldsymbol{\beta}' \mathbf{x}_{it} + \sum_{t=1}^{T} \mathbf{a}_t' \mathbf{x}_{it}) \right] \qquad (8.6)$$

- One key advantage of the CRE estimator: allows for an explicit test for correlation between the individual specific effect and the explanatory variables.
- Drawbacks of the CRE estimator: requires us to impose a good deal of structure on the relationship between α_i and $\mathbf{x}_{it}$ and restricts us to including only time-varying variables in $\mathbf{x}_{it}$.
- Including time invariant variables in $\mathbf{x}_{it}$ in effect induces perfect collinearity since the values of the leads and the lags of these variables will be the same within each cross-sectional unit.
- Its drawbacks aside, this estimator is part of a class of GMM estimators that give substantial efficiency gains over pooled probit estimators, which ignore individual specific effects.
- For details of the estimation procedure for the CRE model, see Hsiao's treatment in *Analysis of Panel Data*, which is more accessible than Chamberlain's original derivation.
- Can be estimated w/ standard software with some more restrictive assumptions; GAUSS code is available for the general model.

8.7.1 CRE Probit Application: PAC contributions and roll call votes

- The standard RE probit does not allow for correlation between voting predispositions and explanatory variables. But presumably, predispositions attract contributions if PACs give to their "friends."
- Tables 3 & 4 report results of the CRE estimation of our model of contributions and votes.
- Figures 2 & 4 plot simulated probabilities based on the CRE estimates.

TABLE 3 Panel Probit Model Results for Voting Behavior on AFL-CIO Roll-Call Votes, 102nd–104th Congresses

	102nd Congress		103rd Congress		104th Congress	
	1st Session	2nd Session	1st Session	2nd Session	1st Session	2nd Session
Variable			***Estimated coefficients****			
ln Labor PAC contributions	–0.006	0.002	–0.012	0.010	0.008	0.012
	(0.004)	(0.011)	(0.009)	(0.006)	(0.001)	(0.009)
ln Corporate PAC contributions	0.031	0.036	0.006	0.010	0.018	0.030
	(0.003)	(0.011)	(0.007)	(0.006)	(0.002)	(0.011)
Unemployment rate	–0.003	–0.008	0.001	–0.006	–0.011	–0.025
	(0.004)	(0.011)	(0.008)	(0.007)	(0.002)	(1.617)
	Tests of no correlation between individual effect and regressors†					
ln Labor PAC contributions	923.348	371.868	342.397	481.322	4,384.273	159.388
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
ln Corporate PAC contributions	170.824	33.342	106.950	32.075	3,986.840	102.342
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
Unemployment rate	112.456	3.222	117.859	31.739	892.259	12.804
	(< 0.001)	(.666)	(< 0.001)	(< 0.001)	(< 0.001)	(.025)
N	315	319	356	332	324	343
T	8	5	7	7	9	5

Note: *Standard errors in parentheses.

† Entries are χ^2_T statistics. *p* values in parentheses.

TABLE 4 Panel Probit Model Results for Voting Behavior on USCC Roll-Call Votes, 102nd–104th Congresses

	102nd Congress		103rd Congress		104th Congress	
	1st Session	2nd Session	1st Session	2nd Session	1st Session	2nd Session
Variable			***Estimated coefficients****			
ln Labor PAC contributions	0.025	–0.005	–0.012	0.043	0.008	0.022
	(0.004)	(0.007)	(0.003)	(0.011)	(0.001)	(0.004)
ln Corporate PAC contributions	–0.030	0.004	0.003	0.010	–0.011	0.015
	(0.005)	(0.006)	(0.004)	(0.009)	(0.001)	(0.004)
Unemployment rate	0.002	–0.010	0.023	0.018	0.004	0.002
	(0.005)	(0.007)	(0.004)	(0.011)	(0.001)	(0.005)
	Tests of no correlation between individual effect and regressors†					
ln Labor PAC contributions	676.705	267.680	119.077	55.355	6,441.995	606.078
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
ln Corporate PAC contributions	459.730	189.141	87.816	37.028	16,767.967	387.557
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
Unemployment rate	106.871	223.512	82.972	15.364	7,427.728	237.892
	(< 0.001)	(< 0.001)	(< 0.001)	(0.031)	(< 0.001)	(< 0.001)
N	302	288	357	344	340	338
T	8	7	7	7	10	8

Note: *Standard errors in parentheses.

† Entries are χ^2_T statistics. *p* values in parentheses.

FIGURE 2 Effect of Labor PAC Contributions on the Probability of Voting in Favor of the AFL-CIO Positions, 104th Congress, 1st Session

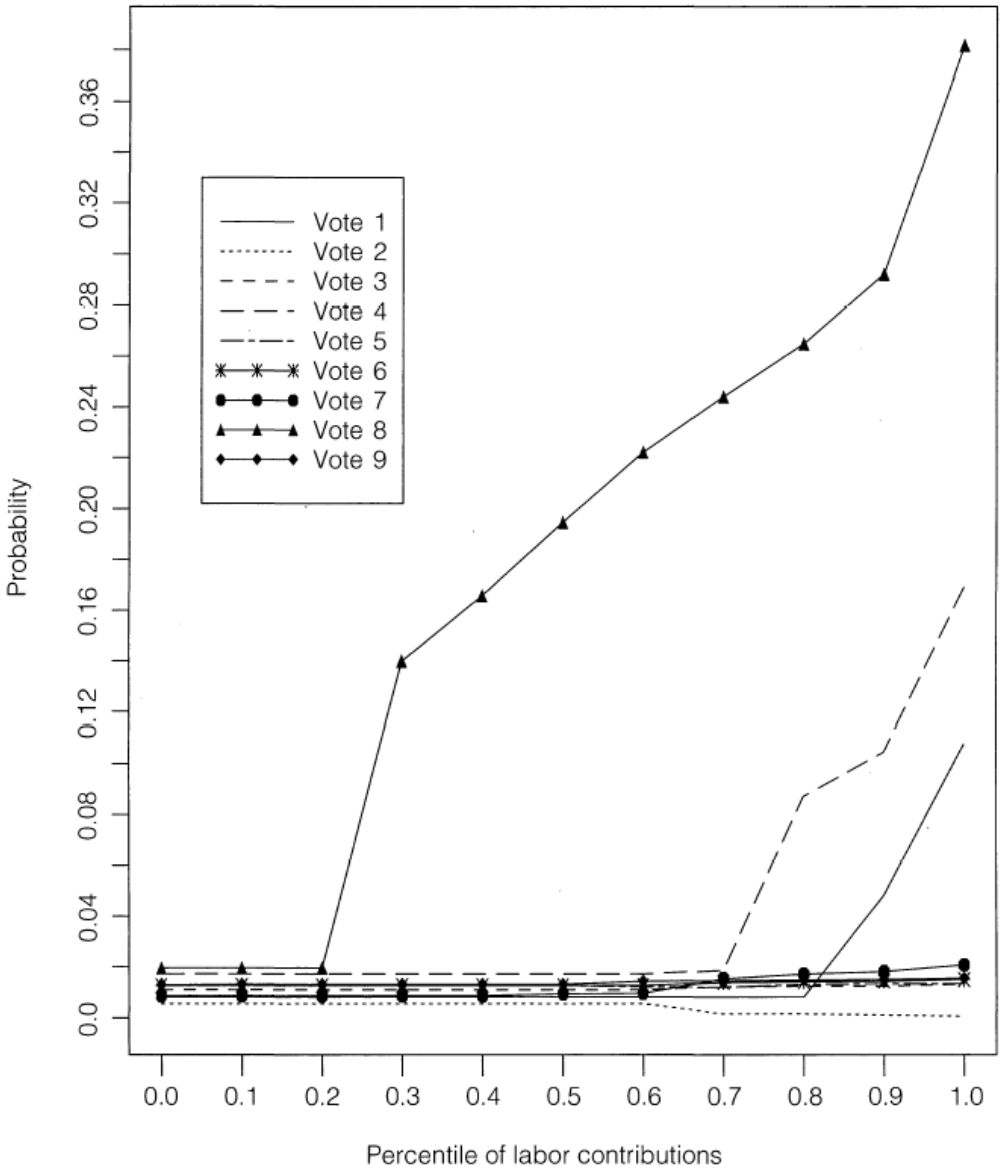
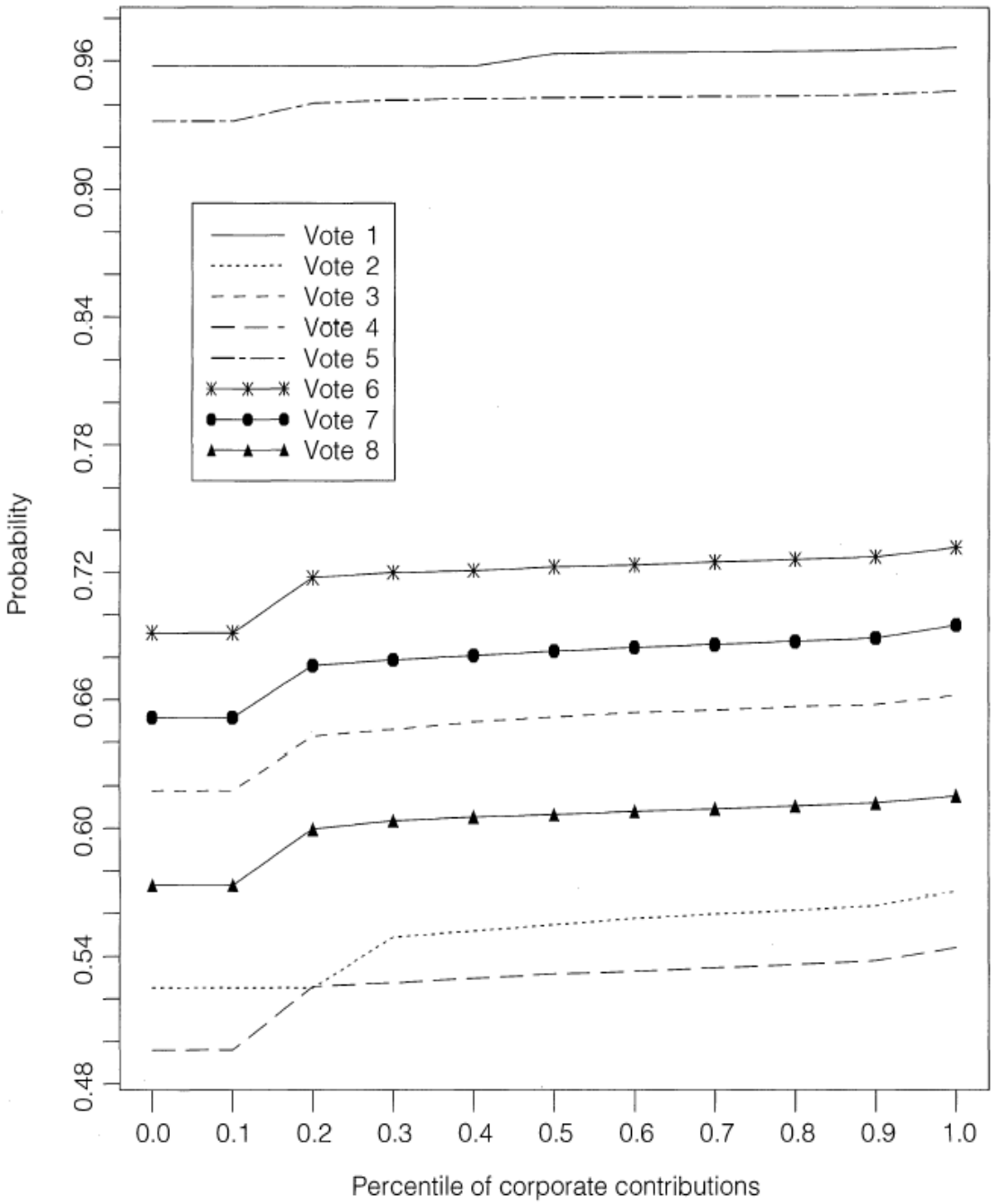


FIGURE 4 Effect of Corporate PAC Contributions on the Probability of Voting in Favor of the USCC Positions, 104th Congress, 2nd Session



8.8 Binary Time-Series Cross-Section (BTSCS) Data

- The methods above are appropriate when N is large and T is small. Beck, Katz, and Tucker ('98 *AJPS*) derive a method for when T is large.
- The method is based on the observation that BTSCS data is identical to grouped duration data. That is, we get to observe whether an event occurred or not only after the end of some discrete period (e.g., a year).
- Thus, we can use duration methods to correct for the problem of temporal dependence.
- Start from the hazard rate for the continuous time Cox proportional hazard model:

$$\lambda(t) = \exp(\boldsymbol{\beta}'\mathbf{x}_{it})\lambda_0(t)$$

- The survival function is given by

$$S(t) = \exp\left(-\int_0^t \lambda(\tau)d\tau\right)$$

- Assuming we get to observe only whether or not an event occurred between time $t_k - 1$ and t_k, we can write

$$\begin{aligned}
\Pr\left(y_{it_k} = 1\right) &= 1 - \exp\left(-\int_{t_k-1}^{t_k} \lambda_i(\tau)d\tau\right) \\
&= 1 - \exp\left(-\int_{t_k-1}^{t_k} \exp(\boldsymbol{\beta}'\mathbf{x}_{it})\lambda_0(t)d\tau\right) \\
&= 1 - \exp\left(-\exp(\boldsymbol{\beta}'\mathbf{x}_{it})\int_{t_k-1}^{t_k} \lambda_0(t)d\tau\right)
\end{aligned}$$

- Let

$$\begin{aligned}
\alpha_{t_k} &= \int_{t_k-1}^{t_k} \lambda_0(t)d\tau \\
\kappa_{t_k} &= \ln(\alpha_{t_k})
\end{aligned}$$

- Then

$$\begin{aligned}
\Pr\left(y_{it_k} = 1\right) &= 1 - \exp\left(-\exp(\boldsymbol{\beta}'\mathbf{x}_{it})\alpha_{t_k}\right) \\
&= 1 - \exp\left(-\exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \kappa_{t_k})\right)
\end{aligned}$$

- This is a binary model with a complimentary log-log (cloglog) link. The cloglog link is identical to a logit link function when the probability of an event is small ($< 25\%$) and extremely similar when the probability of an event is moderate ($< 50\%$).

- For ease of application then, Beck, Katz, and Tucker recommend using the logistic analogue

$$\Pr\left(y_{it}=1|\mathbf{x}_{it}\right)=\frac{1}{1+\exp(-(\boldsymbol{\beta}'\mathbf{x}_{it}+\kappa_{t-t0}))}$$

  where κ_{t-t0} is a dummy variable marking the length of the sequence of zeros that precede the current observation. For example,

t	1	2	3	4	5	6	7	8	9
y	0	0	0	1	0	1	1	0	0
κ	κ_1	κ_2	κ_3	κ_4	κ_1	κ_2	κ_1	κ_1	κ_2

- The intuition behind why ordinary logit is inadequate for BTSCS data is that it doesn't allow for a nonconstant baseline hazard.

- Including the κ dummies allows duration dependence by allowing for a time-varying baseline hazard.

- To see how the κ dummies are interpretable as baseline probabilities or hazards, note

$$\Pr\left(y_{it}=1|\mathbf{x}_{it}=0,t0\right)=\frac{1}{1+\exp(-\kappa_{t-t0})}$$

- The κ dummies are essentially time fixed effects that account for duration dependence. Thus when we estimate the model we need to create a matrix of dummies and concatenate it with the matrix of explanatory variables. For the example given above, this matrix would look like

$$\mathbf{K}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

  Note there are 4 columns because the longest spell is 4 periods long.

$$\Pr\left(y_{it} = 1 | \mathbf{x}_{it}\right) = \frac{1}{1 + \exp(-(\boldsymbol{\beta}' \mathbf{x}_{it} + \kappa_{t-t0}))}$$

8.9 Generalized Estimating Equations (GEEs)

- This class of models allows us to account for unobserved correlation among observations, without including unit-specific effects.
- GEEs relax assumptions about the independence of observations: observations are grouped into clusters and then parameters are estimated to model the correlations among observations in the cluster.
- This class of models is derived from the Generalized Linear Model (GLM) approach.
  - ➢ Models in which one specifies the "link" function $E(Y_i) = \mu_i = h(\boldsymbol{\beta}'\mathbf{x}_i)$ and the relationship between the mean and the variance (e.g. $\mathbf{V}_i = g(\mu_i)$).
  - ➢ Standard GLMs obtain estimates by solving the "score equations":

$$\mathbf{U}_k(\beta) = \sum_{i=1}^{N} \mathbf{D}_i \mathbf{V}_i^{-1}(Y_i - \mu_i) = 0 \tag{8.7}$$

where $\mathbf{D}_i = \frac{\partial \mu_i}{\partial \beta}$ and $\mathbf{V}_i$ is the variance matrix.

8.9.1 GLMs for Correlated Data

- Consider $E(Y_{it}) = \mu_{it} = g(\boldsymbol{\beta}'\mathbf{x}_{it})$, $T > 1$.
- Must make some provision for dependence within i, across t.
- Specify the conditional within-unit or within-cluster correlation:
  - ➢ Define the "working" $T \times T$ correlation matrix $\mathbf{R}_i(\alpha)$ as a function of α.
  - ➢ Structure (but not the elements) of $\mathbf{R}_i(\alpha)$ determined by the investigator.
  - ➢ Label it as "working" b/c we do not expect it to be correctly specified—if not, still get consistent estimates of parameters we care about.
  - ➢ Then redefine the variance matrix in Equation 8.7 as:

$$\mathbf{V}_i = \frac{(\mathbf{A}_i)^{\frac{1}{2}}\mathbf{R}_i(\alpha)(\mathbf{A}_i)^{\frac{1}{2}}}{\phi} \tag{8.8}$$

    where the $\mathbf{A}_i$ are $T \times T$ diagonal matrices w/ $g(\mu_{it})$ along the diagonal and ϕ is a scale parameter typically treated as nuisance.

- Intuition:
  - ➢ Choose β so that μ_{it} is "close" to Y_{it} on average.
  - ➢ Optimally weight each residual $(Y_{it} - \mu_{it})$ by the inverse of $\text{cov}(Y_{it})$.

8.9.2 Options for specifying within-cluster correlation

- "Working Independence": $\mathbf{R}_i(\alpha) = \mathrm{I}$
  - ➢ No within-unit correlation (completely pooling observations).
  - ➢ Just like standard logit/probit.
- "Exchangeable": $\mathbf{R}_i(\alpha) = \{\rho\}_{ts}$, $t \neq s$
  - ➢ Within-unit correlation is the same across units.
  - ➢ Similar to "random effects," in the sense that we're assuming constant within-unit marginal covariance; but do not need orthogonality assumption of RE.
- Autoregressive : $\mathbf{R}_i(\alpha) = \{\rho^{|t-s|}\}_{ts}$
  - ➢ Here, AR(1).
  - ➢ Correlation decays over "time."
  - ➢ Can also do "banded" / "stationary" correlations.
- "Unstructured" : $\mathbf{R}_i(\alpha) = \{\alpha_{ts}\}_{ts}, t \neq s$
  - ➢ $\boldsymbol{\alpha}$ is a $T \times T$ matrix.
  - ➢ Allows $\frac{T(T-1)}{2}$ unique pairwise correlations.
  - ➢ Very flexible, but can be hard to estimate, esp. w/ large T.

8.9.3 "Robust" Standard Errors

- $\hat{\beta}_{GEE}$ is robust to misspecification of $\mathbf{R}_i(\alpha)$.
- $\widehat{\text{var}(\hat{\beta}_{GEE})}$ is not.
- Can compute "sandwich" estimator: $\widehat{\text{var}(\hat{\beta}_{GEE})} =$

$$N\left(\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{S}}_i\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)\left(\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1} \tag{8.9}$$

  where $\hat{\mathbf{S}}_i = (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$.
- Similar to the Huber/White estimator.
- One drawback of GEE is that it does not produce standard errors for elements of $\mathbf{R}_i(\alpha) \Rightarrow$ we should be cautious about drawing inferences about these parameters (treated as "nuisance").
- A number of software packages can estimate GEEs: `xtgee` in Stata; `proc genmod` w/ repeated option in SAS; `gee, geepack` in S-Plus/R; `GEE` in Riemann Library for `GAUSS`.

8.9.4 GEE2

- Other options exist for estimating the $m = \frac{T(T-1)}{2}$ elements of $\mathbf{R}_i(\alpha)$.
- Can do this with a separate estimating equation:

$$\mathbf{U}_{\mathrm{m}}(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \mathbf{E}_i' \mathbf{W}_i^{-1} (\mathbf{Z}_i - \boldsymbol{\eta}_i) \tag{8.10}$$

  where

  - ➢ $\mathbf{Z}_i' = (Z_{i12}, Z_{i13}, \ldots, Z_{i1T}, Z_{i23}, \ldots, Z_{iT,(T-1)})$; the $\frac{T(T-1)}{2}$ "observed" sample pairwise correlation.
  - ➢ $\boldsymbol{\eta}_i$ is the column vector of the expected values of the pairwise intracluster correlation for observation i.
  - ➢ $\mathbf{E}_i = \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\alpha}_i}$
  - ➢ $\mathbf{W}_i$ is a square diagonal matrix of rank $\frac{T(T-1)}{2}$ containing the variances and covariances of the $\mathbf{Z}_i$s.

- Can be estimated either separately from $\mathbf{U}_k(\boldsymbol{\beta})$, assuming $\mathrm{cov}[U_k(\boldsymbol{\beta}), U_m(\alpha)] = 0$, or allowing the two to covary.
- Drawbacks: Requires correct specification of $\mathbf{R}_i(\alpha)$ for consistent estimates of $\hat{\boldsymbol{\beta}}$.

8.9.5 Application: Panel Decision-making on the Court of Appeals

- Question of interest: do gender and race affect the decisions of judges on the U.S. Court of Appeals?
- Almost all cases are decided by (more or less) randomly selected panels of three judges.
- "Norm of unanimity" suggests there will be significant correlation among judges on the same panel.
- GEE approach is useful for accounting for correlation, treating each three judge panel as a cluster.
- Farhang and Wawro (*JLEO* '04) use this approach to examine decisions in employment discrimination cases.

Table 8.1: GEE analysis of judges' votes in Appeals Court decisions

	Logit		GEE	
Variable	Coefficient	Std. Err.	Coefficient	Std. Err.
Judge and Panel Level Variables				
Intercept	1.126	0.362	1.226	0.613
Gender	0.877	0.269	0.791	0.276
One female colleague (female judge)	−0.373	0.454	−0.398	0.594
One female colleague (male judge)	0.800	0.180	0.789	0.274
Two female colleagues (male judge)	−0.337	0.575	−0.362	0.595
Race	−0.329	0.363	−0.212	0.359
One nonwhite colleague (nonwhite judge)	−0.033	0.558	−0.458	0.764
One nonwhite colleague (white judge)	−0.136	0.236	−0.151	0.357
Two nonwhite colleagues (white judge)	−0.443	0.720	−0.859	1.081
NOMINATE score	−0.570	0.192	−0.643	0.176
Panel colleagues' NOMINATE scores	−1.098	0.243	−1.100	0.342
Author	0.122	0.188	−0.023	0.040
Gender × author	−0.174	0.421	0.054	0.120
Race × author	0.345	0.549	0.069	0.117
NOMINATE score × author	−0.173	0.338	0.069	0.082
Case Specific Variables				
Race discrimination	0.285	0.197	0.294	0.310
Gender discrimination	−0.190	0.185	−0.171	0.308
Harassment	0.761	0.206	0.717	0.354
Age discrimination	−0.062	0.190	−0.047	0.330
Religious discrimination	−0.752	0.624	−0.715	1.138
Nationality discrimination	0.069	0.468	0.045	0.784
Reverse gender discrimination	0.576	0.391	0.487	0.716
Reverse race discrimination	−1.282	0.609	−1.388	1.016
Government defendant	0.139	0.159	0.091	0.268
EEOC plaintiff	0.150	0.620	−0.083	1.423
Plaintiff appeal	−1.597	0.235	−1.570	0.390
Both appeal	0.578	0.282	0.586	0.486
Posture	−0.079	0.207	−0.093	0.316
Circuit Level Variables				
1st Circuit dummy	−1.048	0.409	−1.102	0.641
2nd Circuit dummy	−0.624	0.373	−0.636	0.650
3rd Circuit dummy	−1.358	0.549	−1.386	0.773
4th Circuit dummy	−1.885	0.578	−1.973	0.839
5th Circuit dummy	−2.053	0.400	−2.038	0.688
6th Circuit dummy	0.093	0.397	−0.023	0.703
7th Circuit dummy	−1.112	0.320	−1.156	0.541
8th Circuit dummy	−0.870	0.326	−0.923	0.564
10th Circuit dummy	−1.315	0.377	−1.344	0.657
11th Circuit dummy	−0.207	0.334	−0.252	0.583
D.C. Circuit dummy	−1.410	0.489	−1.350	0.794
$\hat{\rho}$	—	—	0.893	—

Note: $N = 1200$.

Section 9

Unbalanced Panels

9.1 Introduction

- Observations can be missing in panel/TSCS data for a number of reasons.
  - ➢ Standard non-response or irregularities in data collection.
  - ➢ By design: rotating panels.
  - ➢ Attrition/accretion.
    - ∗ Individuals can't be located; firms cease to exist.
    - ∗ New firms enter the market; longer time series for some countries.
- The big question is whether or not the missingness is due to an "ignorable" or "nonignorable" selection rule (Verbeek and Nijman in *Econometrics of Panel Data*, '92).
- Ignorability: reason for missingness is unrelated or random to the main DGP of interest.

9.2 Ignorable selection rules

- Consider the simple variance components model:

$$y_{it} = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it} + v_{it} \tag{9.1}$$

where $\alpha_i \sim iidN(0, \sigma_\alpha^2)$, $u_{it} \sim iidN(0, \sigma_u^2)$; assume independence b/t α_i, u_{it}, and $\boldsymbol{\beta}'\mathbf{x}_{it}$, and let $u_{it} = \alpha_i + v_{it}$.

- Let $d_{it} = 0$ if y_{it} (and perhaps elements of $\mathbf{x}_{it}$) are missing; $d_{it} = 1$ if observed.
- Have ignorability (of order 1) for $\boldsymbol{\beta}$ if $E[\boldsymbol{\alpha} + \mathbf{v}_i | \mathbf{d}_i] = \mathbf{0} \ \forall \ i$.
  - ➢ GLS is consistent for unbalanced panel and balanced subpanel if $N \to \infty$;
  - ➢ FE is consistent for unbalanced panel and balanced subpanel if $N \to \infty$ and $E[\tilde{\mathbf{u}}_i | \mathbf{d}_i] = \mathbf{0}$ where $\tilde{\mathbf{u}}_i' = (\tilde{u}_{i1}, \ldots, \tilde{u}_{iT})$ and $\tilde{u}_{it} = u_{it} - \bar{u}_i$.
- W/ ignorability, could just force panels to be balanced by deleting cross-sectional units that have missing obs.; but that's inefficient.
- If we use all available data, then we essentially face an accounting problem in computing estimates: need to keep track of how many obs. we have for each cross-sectional unit.
- Consider the simple case of 2 cross-sectional units observed for different lengths.
- Let n_1 be the shorter time series observed for $i = 1$ and n_2 be the additional obs. we have for $i = 2$.
- Stack observations:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} \tag{9.2}$$

  where $\mathbf{y}_1$ is $n_1 \times 1$, $\mathbf{y}_2$ is $(n_1 + n_2) \times 1$, etc.

- The var-cov matrix is

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_u^2 \mathbf{I}_{n_1} + \sigma_\alpha^2 \mathbf{J}_{n_1 n_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_u^2 \mathbf{I}_{n_1} + \sigma_\alpha^2 \mathbf{J}_{n_1 n_1} & \sigma_\alpha^2 \mathbf{J}_{n_1 n_2} \\ \mathbf{0} & \sigma_\alpha^2 \mathbf{J}_{n_2 n_1} & \sigma_u^2 \mathbf{I}_{n_2} + \sigma_\alpha^2 \mathbf{J}_{n_2 n_2} \end{bmatrix} \tag{9.3}$$

  where $\mathbf{J}_{n_i n_j}$ is an $n_i \times n_j$ matrix of 1s.

- All non-zero off-diagonal elements of $\mathbf{\Omega}$ are equal to σ_α^2.

- Let $T_j = \sum_{i=1}^{j} n_i$ for $j = 1, 2$. Then $\mathbf{\Omega}$ is block diagonal where the jth block is

$$\mathbf{\Omega}_j = (T_j \sigma_\alpha^2 + \sigma_u^2) \bar{\mathbf{J}}_{T_j} + \sigma_u^2 \mathbf{E}_{T_j} \tag{9.4}$$

  where $\bar{\mathbf{J}}_{T_j} = \mathbf{J}_{T_j}/T_j$ and $\mathbf{E}_{T_j} = \mathbf{I}_{T_j} - \bar{\mathbf{J}}_{T_j}$.

- For example, for $T = 3$ w/ the shorter times series being only one period, we have:

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_u^2 + \sigma_\alpha^2 & 0 & \mathbf{0} \\ \mathbf{0} & \sigma_u^2 + \sigma_\alpha^2 & \sigma_\alpha^2 \begin{bmatrix} 1 & 1 \end{bmatrix} \\ \mathbf{0} & \sigma_\alpha^2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \sigma_u^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \sigma_\alpha^2 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_u^2 + \sigma_\alpha^2 & 0 & 0 & 0 \\ 0 & \sigma_u^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 \\ 0 & \sigma_\alpha^2 & \sigma_u^2 + \sigma_\alpha^2 & \sigma_\alpha^2 \\ 0 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_u^2 + \sigma_\alpha^2 \end{bmatrix}$$

- Let $w_j^2 = T_j\sigma_\alpha^2 + \sigma_v^2$.
- Using some math tricks we get

$$\begin{aligned}\sigma_v\boldsymbol{\Omega}_j^{-1/2} &= \frac{\sigma_v}{w_j}\bar{\mathbf{J}}_{T_j} + \mathbf{E}_{T_j} \\ &= \mathbf{I}_{T_j} + \theta_j\bar{\mathbf{J}}_{T_j}\end{aligned}$$

  where $\theta_j = 1 - \sigma_v/w_j$ (note that θ_j varies across j and is a function of T_j).
- We can use this as weights to obtain the GLS estimator as a weighted least squares estimator, where the weights depend on the number of observations for each cross-sectional unit (i.e., T_j).
- Generally, we could write down the variance components model for unbalanced panels as before, but with $t = 1, \ldots, T_i$.
- To compute the within estimator, we modify our standard $\mathbf{Q}$ matrix to be diag($\mathbf{E}_{T_i}$) and then transform the equation (stacking all cross-sectional units and time periods) by pre-multiplying by $\mathbf{Q}$.
- Recall that in the balanced case we pre-multiplied each cross-sectional unit by $\mathbf{Q} = \mathbf{I}_t - \frac{1}{T}\boldsymbol{\iota}\boldsymbol{\iota}'$, which worked b/c we had the same T for each cross-section.
- For the unbalanced case we're pre-multiplying each cross-sectional unit by its own sweep matrix that depends on the length of T for a given unit.

- For the error component derivation, write the regression in vector form:

$$\begin{aligned} \mathbf{y} &= \mu\boldsymbol{\iota}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \\ &= \mathbf{Z}\boldsymbol{\delta} + \mathbf{v} \end{aligned} \tag{9.5}$$

  where

  - ➢ $\mathbf{u} = \mathbf{Z}_\alpha\boldsymbol{\alpha} + \mathbf{v}$,
  - ➢ $n = \sum T_i$,
  - ➢ $\mathbf{y}$ is $n \times 1$; $\mathbf{Z} = (\boldsymbol{\iota}_n, \mathbf{X})$ and is $n \times K$; $\mathbf{Z}_\alpha = \text{diag}(\boldsymbol{\iota}_{T_i})$; $\boldsymbol{\iota}_{T_i}$ is a vector of 1s of dimension T_i;
  - ➢ $\boldsymbol{\delta}' = (\mu, \boldsymbol{\beta}')$.
  - ➢ $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)'$;
    $\mathbf{v} = (v_{11}, \ldots, v_{1T_1}, v_{21}, \ldots, v_{2T_2}, v_{N1}, \ldots, v_{NT_N})'$.

- The GLS estimate is

$$\hat{\boldsymbol{\delta}}_{\text{GLS}} = (\mathbf{Z}'\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Omega}^{-1}\mathbf{y} \tag{9.6}$$

  where $\boldsymbol{\Omega} = \sigma_v^2\boldsymbol{\Sigma}$,

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathbf{I}_n + \rho\mathbf{Z}_\alpha\mathbf{Z}_\alpha' \\ &= \text{diag}(\mathbf{E}_{T_i}) + \text{diag}[(1 + \rho T_i)\bar{\mathbf{J}}_{T_i}], \end{aligned} \tag{9.7}$$

  and $\rho = \sigma_\alpha^2/\sigma_v^2$.

- $(1 + \rho T_i) = w_i^2/\sigma_v^2$ where $w_i^2 = T_i\sigma_\alpha^2 + \sigma_v^2$ (see eq. 9.4).

- Thus, GLS estimates can be obtained by running OLS on the original equation after we have pre-multiplied it by

$$\sigma_v\boldsymbol{\Omega}^{-1/2} = \text{diag}(\mathbf{E}_{T_i}) + \text{diag}[(\sigma_v/w_i)\bar{\mathbf{J}}_{T_i}] \tag{9.8}$$

  (see eq. 9.2).

- In other words,
$$\hat{\boldsymbol{\delta}}_{\text{GLS}} = (\mathbf{Z}^{*\prime}\mathbf{Z}^{*})^{-1}\mathbf{Z}^{*\prime}\mathbf{y}^{*} \tag{9.9}$$
where $\mathbf{Z}^{*} = \sigma_v \boldsymbol{\Omega}^{-1/2}\mathbf{Z}$ and $\mathbf{y}^{*} = \sigma_v \boldsymbol{\Omega}^{-1/2}\mathbf{y}$.

9.3 Nonignorable selection

- If the selection is nonignorable, then we are faced w/ a more difficult problem.
- One way to fix this is to try to incorporate the selection rule in our model.

9.3.1 Review of selection in the cross-sectional case

- Latent structure:
$$\begin{aligned} d_i^{*} &= \boldsymbol{\delta}'\mathbf{w}_i + \epsilon_i \\ y_i^{*} &= \boldsymbol{\beta}'\mathbf{x}_i + u_i \end{aligned}$$
- Observed structure:
$$d_i = \begin{cases} 1 & \text{if } d_i^{*} > 0 \\ 0 & \text{if } d_i^{*} \leq 0 \end{cases} \qquad y_i = \begin{cases} y_i^{*} & \text{if } d_i = 1 \\ ? & \text{otherwise} \end{cases}$$
- Assume distribution for the disturbances:
$$\begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon u} \\ \sigma_{\epsilon u} & \sigma_u^2 \end{bmatrix}\right)$$

- That is, we assume that (ϵ_i, u_i) have a bivariate normal distribution:

$$f(\epsilon_i, u_i) = \frac{1}{2\pi\sigma_\epsilon\sigma_u\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\epsilon_i}{\sigma_\epsilon}\right)^2 - 2\rho\frac{\epsilon_i u_i}{\sigma_\epsilon\sigma_u} + \left(\frac{u_i}{\sigma_u}\right)^2\right]\right\}$$

  where $\rho = \frac{\sigma_{\epsilon u}}{\sigma_\epsilon\sigma_u}$.

- What is the likelihood function for this model?

  Case 1: $d_i = 0$, so $y_i = 0$.

  Individual's contribution to the likelihood is $\Pr(d_i = 0|\mathbf{w}_i) = 1 - \Phi(\boldsymbol{\delta}'\mathbf{w}_i/\sigma_\epsilon)$.

  Case 2: $d_i = 1$ and we get to see y_i.

  To figure out what an individual's contribution to the likelihood is in this case will be we need to think about the distribution of u_i given $d_i = 1$:

$$\int_c^\infty f(\epsilon_i, u_i)d\epsilon_i \tag{9.10}$$

  where c is the threshold ϵ_i must exceed to observe y_{2i} (i.e., $c = -\boldsymbol{\delta}'\mathbf{w}_i$).

  Trick: a joint distribution can be written as a conditional distribution times a marginal distribution:

$$f(\epsilon_i, u_i) = f(u_i)f(\epsilon_i|u_i).$$

Then (9.10) can be written as

$$\begin{aligned}\int_c^\infty f(\epsilon_i, u_i)d\epsilon_i &= \int_c^\infty f(u_i)f(\epsilon_i|u_i)d\epsilon_i \\ &= f(u_i)\int_c^\infty f(\epsilon_i|u_i)d\epsilon_i \\ &= f(u_i)\Pr(\epsilon_i > c|u_i)\end{aligned}$$

Given our assumptions, $\epsilon_i|u_i \sim N\left(\frac{\sigma_{\epsilon u}}{\sigma_u^2}u_i, (1-\rho^2)\sigma_\epsilon^2\right)$.

What is $\Pr(\epsilon_i > c|u_i)$?

$$\frac{\epsilon_i - \frac{\sigma_{\epsilon u}}{\sigma_u^2}u_i}{(1-\rho^2)^{1/2}\sigma_\epsilon} > \frac{c - \frac{\sigma_{\epsilon u}}{\sigma_u^2}u_i}{(1-\rho^2)^{1/2}\sigma_\epsilon}$$

Let $c = -\boldsymbol{\delta}'\mathbf{w}_i$ and $u_i = y_i - \boldsymbol{\beta}'\mathbf{x}_i$. Then

$$\frac{c - \frac{\sigma_{\epsilon u}}{\sigma_u^2}u_i}{(1-\rho^2)^{1/2}\sigma_\epsilon} = \frac{-\boldsymbol{\delta}'\mathbf{w}_i - \frac{\sigma_{\epsilon u}}{\sigma_u^2}(y_i - \boldsymbol{\beta}'\mathbf{x}_i)}{(1-\rho^2)^{1/2}\sigma_\epsilon}$$

- Thus an individual's contribution to the likelihood if $d_i = 1$ is

$$\frac{1}{\sigma_u}\phi\left(\frac{y_i - \boldsymbol{\beta}'\mathbf{x}_i}{\sigma_u}\right)\left[1 - \Phi\left(\frac{-\boldsymbol{\delta}'\mathbf{w}_i - \frac{\sigma_{\epsilon u}}{\sigma_u^2}(y_i - \boldsymbol{\beta}'\mathbf{x}_i)}{(1-\rho^2)^{1/2}\sigma_\epsilon}\right)\right]$$

- The log likelihood function for this model is

$$\sum_{d_i=0} \ln\left[1-\Phi(\boldsymbol{\delta}'\mathbf{w}_i/\sigma_\epsilon)\right] + \sum_{d_i=1} \ln\left(\frac{1}{\sigma_u}\phi\left(\frac{y_i-\boldsymbol{\beta}'\mathbf{x}_i}{\sigma_u}\right)\right)$$
$$+ \sum_{d_i=1} \ln\left[1-\Phi\left(\frac{-\boldsymbol{\delta}'\mathbf{w}_i - \frac{\sigma_{\epsilon u}}{\sigma_u^2}(y_i-\boldsymbol{\beta}'\mathbf{x}_i)}{(1-\rho^2)^{1/2}\sigma_\epsilon}\right)\right]$$

- As you can well imagine, this log likelihood is not the easiest thing to maximize.

  ➢ Life would be easier if we had starting values that put us close to the maximum.

  ➢ **Heckman's two-step method** gives us such starting values, but this method is often used instead of ML.

- Consider this model in a regression setting:

$$E[d_i|\mathbf{w}_i] = \Pr(d_i=1|\mathbf{w}_i) = \Phi(\boldsymbol{\delta}'\mathbf{w}_i/\sigma_\epsilon)$$

$$\begin{aligned} E[y_i|\mathbf{x}_i, d_i=1] &= E[\boldsymbol{\beta}'\mathbf{x}_i + u_i|\mathbf{x}_i, d_i=1] \\ &= \boldsymbol{\beta}'\mathbf{x}_i + E[u_i|\mathbf{x}_i, d_i=1] \\ &= \boldsymbol{\beta}'\mathbf{x}_i + E[u_i|\epsilon_i > -\boldsymbol{\delta}'\mathbf{w}_i] \end{aligned}$$

- How might this be problematic for using OLS?
  - ➢ if $E[u_i|\epsilon_i > -\boldsymbol{\delta}'\mathbf{w}_i] = 0$, no bias.
  - ➢ if $E[u_i|\epsilon_i > -\boldsymbol{\delta}'\mathbf{w}_i] = a$ where a is any nonzero constant, then we get bias in the intercept but not the slope coefficients.
  - ➢ if $E[u_i|\epsilon_i > -\boldsymbol{\delta}'\mathbf{w}_i] = f(\boldsymbol{\beta}'\mathbf{x}_i)$, then we get inconsistency in the intercept and the slope coefficients.

- We can use the **law of iterated expectations** to figure out what $E[u_i|\epsilon_i > -\boldsymbol{\delta}'\mathbf{w}_i]$ is:

$$E[y] = E_x[E[y|x]]$$

$$E[u_i|\epsilon_i > -\boldsymbol{\delta}'\mathbf{w}_i] = E_\epsilon\left\{E[u_i|\epsilon_i > -\boldsymbol{\delta}'\mathbf{w}_i, \epsilon_i]|\epsilon_i > -\boldsymbol{\delta}'\mathbf{w}_i]\right\}$$

- Take the inner expectation:

$$\begin{aligned} E[u_i|\epsilon_i > -\boldsymbol{\delta}'\mathbf{w}_i, \epsilon_i] &= E[u_i|\epsilon_i] \\ &= \tau\epsilon_i \end{aligned}$$

  where

$$\tau = \frac{\text{cov}(\epsilon_i, u_i)}{\text{var}(\epsilon_i)} = \frac{\sigma_{\epsilon u}}{\sigma_\epsilon^2}$$

- The outer expectation then becomes

$$\begin{aligned} E[\tau\epsilon_i|\epsilon_i > -\boldsymbol{\delta}'\mathbf{w}_i] &= \tau E[\epsilon_i|\epsilon_i > -\boldsymbol{\delta}'\mathbf{w}_i] \\ &= \tau\sigma_\epsilon E\left[\frac{\epsilon_i}{\sigma_\epsilon}|\frac{\epsilon_i}{\sigma_\epsilon} > \frac{-\boldsymbol{\delta}'\mathbf{w}_i}{\sigma_\epsilon}\right] \\ &= \tau\sigma_\epsilon\frac{\phi(\boldsymbol{\delta}'\mathbf{w}_i/\sigma_\epsilon)}{\Phi(\boldsymbol{\delta}'\mathbf{w}_i/\sigma_\epsilon)} \end{aligned}$$

- This follows from the behavior of moments of the truncated normal distribution.

  ➢ $\Phi(\boldsymbol{\delta}'\mathbf{w}_i/\sigma_\epsilon)$ in the denominator is essentially a weight that ensures that the distribution will integrate to 1.

- Putting this altogether gives

$$E[y|\mathbf{x}_i, d_i = 1] = \boldsymbol{\beta}'\mathbf{x}_i + \rho\sigma_u\frac{\phi(\boldsymbol{\delta}'\mathbf{w}_i/\sigma_\epsilon)}{\Phi(\boldsymbol{\delta}'\mathbf{w}_i/\sigma_\epsilon)}$$

9.4 Selection for repeated observations data

- Let's derive this in the error components framework:

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + v_{it} \tag{9.11}$$

  where $v_{it} = \alpha_i + u_{it}$.

- Suppose we have $T = 2$ w/ attrition in the 2nd period, and let $d_i = 1$ if y_{i2} is observed, and is 0 otherwise.

- Latent variable set-up:

$$d_i^* = \gamma y_{i2} + \boldsymbol{\theta}'\mathbf{x}_{i2} + \boldsymbol{\delta}'\mathbf{w}_i + \epsilon_i^* \geq 0 \tag{9.12}$$

(assume joint normal for $\mathbf{v}_i$ and ϵ^*).

- Substitute in for y_{i2} and rearrange:

$$\begin{aligned} d_i^* &= (\gamma\boldsymbol{\beta}' + \boldsymbol{\theta}')\mathbf{x}_{i2} + \boldsymbol{\delta}'\mathbf{w}_i + \gamma v_{i2}\epsilon_i^* \\ &= \boldsymbol{\pi}'\mathbf{x}_{i2} + \boldsymbol{\delta}'\mathbf{w}_i + \epsilon_i \\ &= \mathbf{a}'\mathbf{R}_i + \epsilon_i \end{aligned} \tag{9.13}$$

where $\epsilon_i = \gamma v_{i2}\epsilon_i^*$, $\mathbf{R}_i = (\mathbf{x}_{i2}' + \boldsymbol{\delta}'\mathbf{w}_i')'$, and $\mathbf{a}' = (\boldsymbol{\pi}', \boldsymbol{\delta}')$

- The probabilities of retention and attrition are:

$$\begin{aligned} \Pr(d_i = 1) &= \Phi(\mathbf{a}'\mathbf{R}_i) \\ \Pr(d_i = 0) &= 1 - \Phi(\mathbf{a}'\mathbf{R}_i) \end{aligned} \tag{9.14}$$

- The condit'l expectation of y_{i2} given that it is observed is

$$E(y_{i2}|\boldsymbol{\beta}'\mathbf{x}_{i2}, \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}'\mathbf{x}_{i2} + E(v_{i2}|\boldsymbol{\beta}'\mathbf{x}_{i2}, \mathbf{w}_i, d_i = 1) \tag{9.15}$$

- With $v_{i2} = \sigma_{2\epsilon}\epsilon + \eta_i$ where $\sigma_{2\epsilon} = \text{cov}[v_{i2}, \epsilon]$ and $\eta_i \perp \epsilon_i$, we get:

$$\begin{aligned} E(v_{i2}|\mathbf{w}_i, d_i = 1) &= \sigma_{2\epsilon}E(\epsilon_i|\mathbf{w}_i, d_i = 1) \\ &= \frac{\sigma_{2\epsilon}}{\Phi(\mathbf{a}'\mathbf{R}_i)} \int_{-\mathbf{a}'\mathbf{R}_i}^{\infty} \epsilon \cdot \frac{1}{\sqrt{2\pi}} e^{-\epsilon^2/2} d\epsilon \\ &= \epsilon_{2\epsilon} \frac{\phi(\mathbf{a}'\mathbf{R}_i)}{\Phi(\mathbf{a}'\mathbf{R}_i)} \end{aligned} \tag{9.16}$$

This is essentially the same as what we did in the cross-sectional case (finding the expectation of a truncated, normally distributed random variable).

- This gives

$$E(y_{i2}|\boldsymbol{\beta}'\mathbf{x}_{i2}, \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}'\mathbf{x}_{i2} + \epsilon_{2\epsilon}\frac{\phi(\mathbf{a}'\mathbf{R}_i)}{\Phi(\mathbf{a}'\mathbf{R}_i)} \tag{9.17}$$

- One way to get estimates is via ML. The joint density of $d_i = 1, y_{i1}$, and y_{i2} is

$$\begin{aligned} f(d_i = 1, y_{i1}, y_{i2}) &= \Pr(d_i = 1|y_{i1}, y_{i2}) f(y_{i1}, y_{i2}) \\ &= \Pr(d_i = 1|y_{i2}) f(y_{i1}, y_{i2}) \\ &= \Phi\left\{ \frac{\mathbf{a}'\mathbf{R}_i + \left(\frac{\sigma_{2\epsilon}}{\sigma_u^2+\sigma_\alpha^2}\right)(y_{i2} - \boldsymbol{\beta}'\mathbf{x}_{i2})}{\left[1 - \frac{\sigma_{2\epsilon}}{\sigma_u^2+\sigma_\alpha^2}\right]^{1/2}} \right\} \\ &\times [2\pi\sigma_u^2(\sigma_u^2 + 2\sigma_\alpha^2)]^{-1/2} \\ &\times \exp\left\{ -\frac{1}{2\sigma_u^2}\left[\sum_{t=1}^{2}(y_{it} - \boldsymbol{\beta}'\mathbf{x}_{it})^2 - \frac{\sigma_\alpha^2}{\sigma_u^2 + 2\sigma_\alpha^2} \right.\right. \\ &\times \left.\left. \left(\sum_{t=1}^{2}(y_{it} - \boldsymbol{\beta}'\mathbf{x}_{it}) \right)^2 \right] \right\}, \end{aligned} \tag{9.18}$$

  ➢ The $\Phi(\cdot)$ term in this equation comes from

$$f(\epsilon_i|v_{i2}) \sim N\left(\frac{\sigma_{2\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} v_{i2}, 1 - \frac{\sigma_{2\epsilon}^2}{\sigma_u^2 + \sigma_\alpha^2} \right) \tag{9.19}$$

  ➢ The remaining parts of the equation come from the joint normal for y_{i1} and y_{i2}.

- An individual who has left the sample contributes the following to the likelihood:

$$
\begin{aligned}
f(d_i = 0, y_{i1}) &= \Pr(d_i = 0 | y_{i1}) f(y_{i1}) \\
&= \left\{ 1 - \Phi \left[\frac{\mathbf{a}'\mathbf{R}_i + \left(\frac{\sigma_{1\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} \right) (y_{i1} - \boldsymbol{\beta}'\mathbf{x}_{i1})}{\left[1 - \frac{\sigma_{1\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} \right]^{1/2}} \right] \right\} \\
&\times [2\pi(\sigma_u^2 + 2\sigma_\alpha^2)]^{-1/2} \\
&\times \exp \left\{ -\frac{1}{2(\sigma_u^2 + \sigma_\alpha^2)} (y_{i1} - \boldsymbol{\beta}'\mathbf{x}_{i1})^2 \right\}
\end{aligned}
\tag{9.20}
$$

where $\sigma_{1\epsilon} = \text{cov}[\epsilon_i, v_{i1}] = \sigma_{2\epsilon} = \sigma_\alpha^2/(\sigma_u^2 + \sigma_\alpha^2)$.

- Order the observations so that we have complete data on the first N_1 of them and the remaining $N - N_1$ represent those who have left the sample after the first period. Then the log likelihood is

$$\begin{aligned}
\ln L = & -N \ln 2\pi - \frac{N_1}{2} \ln \sigma_u^2 - \frac{N_1}{2} \ln(\sigma_u^2 + 2\sigma_\alpha^2) \\
& - \frac{N - N_1}{2} \ln(\sigma_u^2 + \sigma_\alpha^2) \\
& - \frac{1}{2\sigma_u^2} \sum_{i=1}^{N_1} \left\{ \sum_{t=1}^{2} (y_{it} - \boldsymbol{\beta}' \mathbf{x}_{it})^2 - \frac{\sigma_\alpha^2}{\sigma_u^2 + 2\sigma_\alpha^2} \left[\sum_{t=1}^{2} (y_{it} - \boldsymbol{\beta}' \mathbf{x}_{it}) \right]^2 \right\} \\
& + \sum_{i=1}^{N_1} \ln \Phi \left\{ \frac{\mathbf{a}' \mathbf{R}_i + \left(\frac{\sigma_{2\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} \right) (y_{i2} - \boldsymbol{\beta}' \mathbf{x}_{i2})}{\left[1 - \frac{\sigma_{2\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} \right]^{1/2}} \right\} \\
& - \frac{1}{2(\sigma_u^2 + \sigma_\alpha^2)} \sum_{i=N_1+1}^{N} (y_{i1} - \boldsymbol{\beta}' \mathbf{x}_{i1})^2 \\
& + \sum_{i=N_1+1}^{N} \ln \left\{ 1 - \Phi \left[\frac{\mathbf{a}' \mathbf{R}_i + \left(\frac{\sigma_{1\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} \right) (y_{i1} - \boldsymbol{\beta}' \mathbf{x}_{i1})}{\left[1 - \frac{\sigma_{1\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} \right]^{1/2}} \right] \right\}
\end{aligned} \tag{9.21}$$

- Note that if $\sigma_{2\epsilon} = 0$ then there is no attrition bias (also means that $\sigma_{1\epsilon} = 0$); get random attrition model.
- Can be generalized for $T > 2$; specify attrition equation for each period.