From C&G(N)LRM to Models for TSCS

I. Notation, most-general (linear) model:

$$y_{it} = \alpha_{it} + \mathbf{x}_{it}'\boldsymbol{\beta}_{it} + \varepsilon_{it};\ \ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma}_{it});\ \ i = 1..N,\ t = 1..T,\ n = NT$$

A. Nothing necessarily changes (all data are TSCS data):

1. If willing assume (or *insofar as*) Gauss-Markov holds:

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it};\ \ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon};\ \ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$$

$$Cov(\mathbf{X}, \boldsymbol{\varepsilon}) = \mathbf{0}$$

$$\left\{\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})\right\}$$

a) Last not nec for OLS (if absent: CLT for dists of ests); nec for ML=OLS

b) Nothing new; this C(N)LRM => OLS=BLUE.

(1) BLUE: Best Linear Unbiased Estimator =>

(2) Coefficient-Estimate (**b**) Properties: unbiased, consistent, efficient

(3) V-Cov(**b**)-Estimate Properties: unbiased, consistent, efficient

2. Similarly, if want relax to: $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon};\ \ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Omega});\ \ Cov(\mathbf{X}, \boldsymbol{\varepsilon}) = \mathbf{0}$

   a) (w/ normality again as above) nothing new: G(N)LRM => GLS=BLUE, FGLS=asymptotically BLUE, where asymptotically BLUE means:

   b) **FGLS Properties** {of **b** & est'd V(**b**)}**:** Consistent, Asymptotically Efficient

3. TSCS=collection of time-series, so all you may know regarding TS models applies (w/approp care to respect breaks b/w units; eg, $y_{2,1}$(t-1)$\neq y_{1,T}$)

B. Thus, departure from C(N)LRM lies in plausibility key assumptions:

1. Parameter Stability:

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \Rightarrow (\alpha_{it}, \boldsymbol{\beta}_{it}) = (\alpha, \boldsymbol{\beta})\ \forall i, t$$

2. Spherical Errors (homoskedasticity+uncorrelated):

$$\text{some } \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\boldsymbol{\Omega}) \text{ more plausible than } \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$$

3. Parameters Vary? Stochastic Component Variance-Covariance Structure Varies? => First Line of Defense, Always & Everywhere: MODEL IT!

II. From the Most-General (& Inestimable) Form Down:

A. Most-General Form:

$$y_{it} = f_{it}\left(\mathbf{x}_{it}, \boldsymbol{\beta}_{it}, \varepsilon_{it}\right);\ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma}_{it});\ i = 1..N,\ t = 1..T,\ n = NT$$

<u>Notes</u>: $\mathbf{x}^1 = 1$; $\boldsymbol{\beta}^1 = \alpha$; $\mathbf{x}$ may contain time-space lags $\mathbf{x}$ or $\mathbf{y}$.

1. Parameters=K+½(NT)^2+½NT *per function*, ***per observation***!

2. MASSIVELY under-identified ⇒ impose structure to reduce parameterization; from where? Theory & Substance (a.k.a., Assumptions & Stories)

B. Virtually always assume:

1. $f_{it}(\bullet) = f(\bullet)\ \forall i,t$: same *f*() relates $\mathbf{X}_{it}$,$\boldsymbol{\beta}_{it}$,$\boldsymbol{\varepsilon}_{it}$ to **y** in all obs; may be stronger than needed; could parameterize changes $f_{it}(\bullet)$, or allow it to vary across but not within groups of obs $\{it\}$, or many other relaxations, constrained by deg free >0.

2. $\boldsymbol{\Sigma}_{it} = \boldsymbol{\Sigma}\ \forall i,t$: each obs draw from distribution with same variance-covariance across obs; may be stronger than needed…

3. Parameters: Still K(NT)+½(NT)^2+½NT per NT obs.=>K+½(NT+1) per *observation*. Still **<u>way</u>**, **<u>way</u>** too many.

C. Next, can assume constant coefficient-vector: $\boldsymbol{\beta}_{it} = \boldsymbol{\beta}\ \forall i,t$

1. Parameters: K/(NT)+½(NT+1) per obs. Still <u>**way**</u> too many.

2. May be stronger than needed, can allow: $\boldsymbol{\beta}_{it} = g\left(\mathbf{z}, \boldsymbol{\gamma}, \eta_{it}\right)$, with # of parameters < NT-K-# free parameters in $\boldsymbol{\Sigma}$.

D. Still must reduce parameterization $\boldsymbol{\Sigma}$ (*see detail of* $\boldsymbol{\Sigma}$ *on next slide*).

1. Fully general var-covar structure not estimable; nothing to learn from history &/or comparison if insist all unique.

2. Plausible/practically-realistic variance-covariance structures:

a) *Sphericity*: from $\sigma^2\boldsymbol{\Omega}$ to $\sigma^2\mathbf{I}$ => from ½(NT)^2+½NT to 1 parameter.

b) *Panel Heteroskedasticity*: from $V\left(\varepsilon_{it}\right) = \sigma_i^2$ => N parameters.

c) *Serial Correlation*: (AR1) $\varepsilon_{it} = \rho\varepsilon_{i,t-1} + \upsilon$ (2 params), or $\varepsilon_{it} = \rho_i\varepsilon_{i,t-1}$ ($N+1$ pars)

d) *Parks-Kmenta*: panel het + each TS unique AR1 + unique $\sigma_{ij} = \sigma_{ji}\ \forall ij$, though common for all T $\Rightarrow$ $2N$+½N(N-1) $\Rightarrow$ needs <u>LOTS</u> of T.

e) *Many other plausible parameterizations…*

$$
\boldsymbol{\Omega} =
\begin{bmatrix}
\omega_{11}^2 & \omega_{11,12} & \omega_{11,13} & \cdots & \omega_{11,1T} & \omega_{11,21} & \omega_{11,22} & \omega_{11,23} & \cdots & \omega_{11,2T} & \omega_{11,31} & \omega_{11,32} & \omega_{11,33} & \cdots & \omega_{11,3T} & & \omega_{11,N1} & \omega_{11,N2} & \omega_{11,N3} & \cdots & \omega_{11,NT} \\
\omega_{12,11} & \omega_{12}^2 & & & \vdots & \omega_{12,21} & \omega_{12,22} & & & \vdots & \omega_{12,31} & \omega_{12,32} & & & \vdots & & \omega_{12,N1} & \omega_{12,N2} & & & \vdots \\
\omega_{13,11} & & \omega_{13}^2 & & \vdots & \omega_{13,21} & & \omega_{13,23} & & \vdots & \omega_{13,31} & & \omega_{13,33} & & \vdots & \cdots & \omega_{13,N1} & & \omega_{13,N3} & & \vdots \\
\vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots & & \vdots & & & \ddots & \vdots \\
\omega_{1T,11} & \cdots & \cdots & \cdots & \omega_{1T}^2 & \omega_{1T,21} & \cdots & \cdots & \cdots & \omega_{1T,2T} & \omega_{1T,31} & \cdots & \cdots & \cdots & \omega_{1T,3T} & & \omega_{1T,N1} & \cdots & \cdots & \cdots & \omega_{1T,NT} \\
\omega_{21,11} & \omega_{21,12} & \omega_{21,13} & \cdots & \omega_{21,1T} & \omega_{21}^2 & \omega_{21,22} & \omega_{21,23} & \cdots & \omega_{21,2T} & & & & & & & & & & & \\
\omega_{22,11} & \omega_{22,12} & & & \vdots & \omega_{21,22} & \omega_{22}^2 & & & \vdots & & & & & & & & & & & \\
\omega_{23,11} & & \omega_{23,13} & & \vdots & \omega_{21,23} & & \omega_{23}^2 & & \vdots & & & & & & & & & \vdots & & \\
\vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots & & & & & & & & & & & \\
\omega_{2T,11} & \cdots & \cdots & \cdots & \omega_{2T,1T} & \omega_{21,2T} & \cdots & \cdots & \cdots & \omega_{2T}^2 & & & & & & & & & & & \\
\omega_{31,11} & \omega_{31,12} & \omega_{31,13} & \cdots & \omega_{31,1T} & & & & & & \omega_{31}^2 & \omega_{31,32} & \omega_{31,33} & \cdots & \omega_{31,3T} & & & & & & \\
\omega_{32,11} & \omega_{32,12} & & & \vdots & & & & & & \omega_{32,31} & \omega_{32}^2 & & & \vdots & & & & & & \\
\omega_{33,11} & & \omega_{33,13} & & \vdots & & & & & & \omega_{33,31} & & \omega_{33}^2 & & \vdots & & & & \vdots & & \\
\vdots & & & \ddots & \vdots & & & & & & \vdots & & & \ddots & \vdots & & & & & & \\
\omega_{3T,11} & \cdots & \cdots & \cdots & \omega_{3T,1T} & & & & & & \omega_{3T,31} & \cdots & \cdots & \cdots & \omega_{3T}^2 & & & & & & \\
 & & \vdots & & & & & & & & & & & & & \ddots & & & \vdots & & \\
\omega_{N1,11} & \omega_{N1,12} & \omega_{N1,13} & \cdots & \omega_{N1,1T} & & & & & & & & & & & & \omega_{N1}^2 & \omega_{N1,N2} & \omega_{N1,N3} & \cdots & \omega_{N1,NT} \\
\omega_{N2,11} & \omega_{N2,12} & & & \vdots & & & & & & & & & & & & \omega_{N2,N1} & \omega_{N2}^2 & & & \vdots \\
\omega_{N3,11} & & \omega_{N3,13} & & \vdots & & & \cdots & & & & & \cdots & & & \cdots & \omega_{N3,N1} & & \omega_{N3}^2 & & \vdots \\
\vdots & & & \ddots & \vdots & & & & & & & & & & & & \vdots & & & \ddots & \vdots \\
\omega_{NT,11} & \cdots & \cdots & \cdots & \omega_{NT,1T} & & & & & & & & & & & & \omega_{NT,N1} & \cdots & \cdots & \cdots & \omega_{NT}^2
\end{bmatrix}
$$

III. From simplest model upward (parsimony principle):

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}, \quad \mathrm{V}\left(\varepsilon_{it}\right) = \sigma^2$$

A. Pool all data and estimate by OLS

1. Advantages:

   a) Gives maximal leverage estimating parameters (i.e., highly efficient);

   b) Consistent w/ general theories;

   c) BLUE, *iff* this right model…

2. Disadvantages:

   a) Might not be right model (i.e., inefficiency & likely bias if wrong model).

   b) What might commonly expect wrong (specification error; omitted-variables):

      (1) Nonsphericity: $\mathrm{V}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{\Omega} \neq \sigma^2 \mathbf{I}$

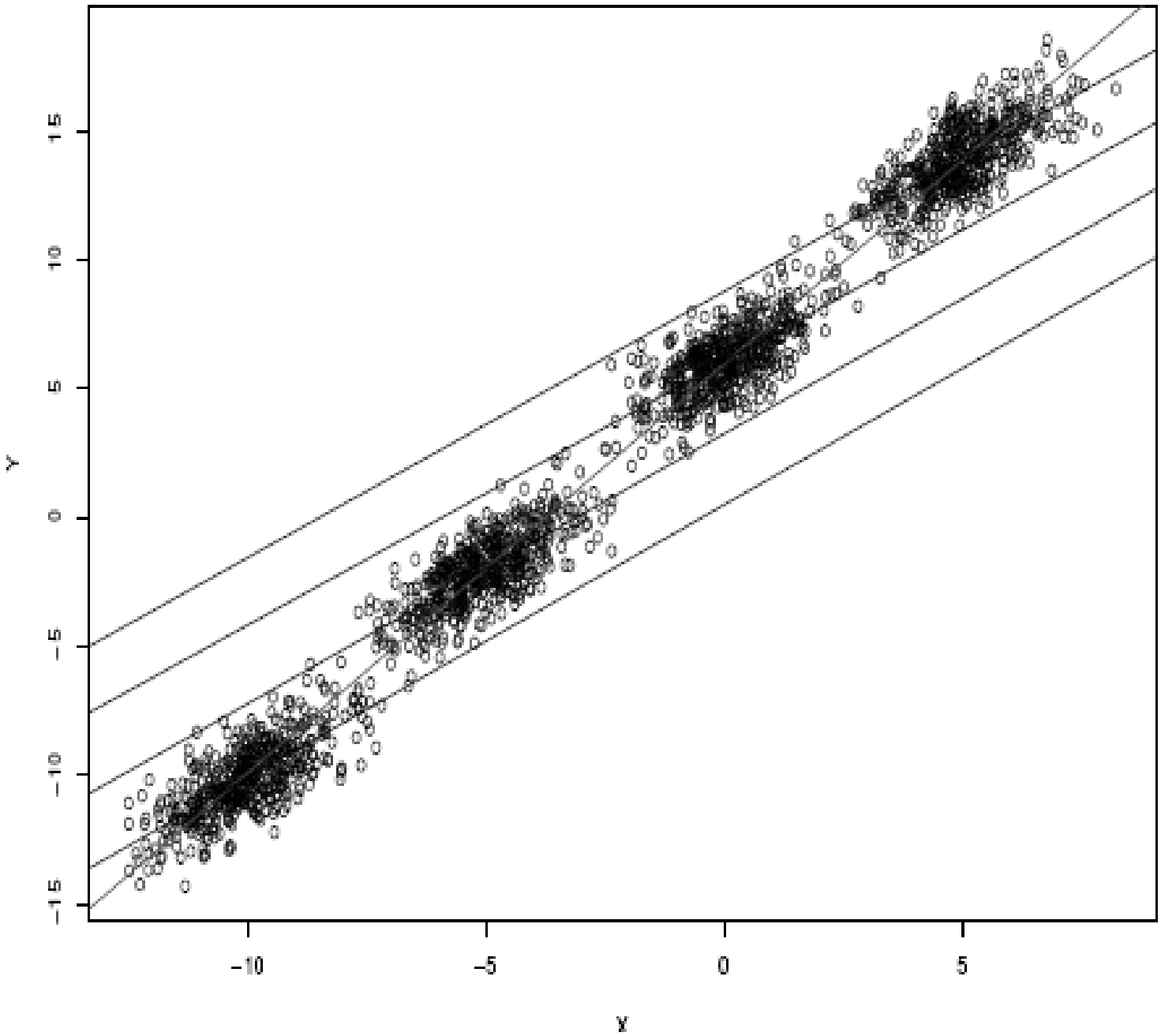
      (2) Unobserved (unmodeled) Unit (e.g., country) Effects: $\alpha_i \neq \alpha$

      (3) Unobserved (unmodeled) Time (or sub-unit) Effects: $\alpha_t \neq \alpha$

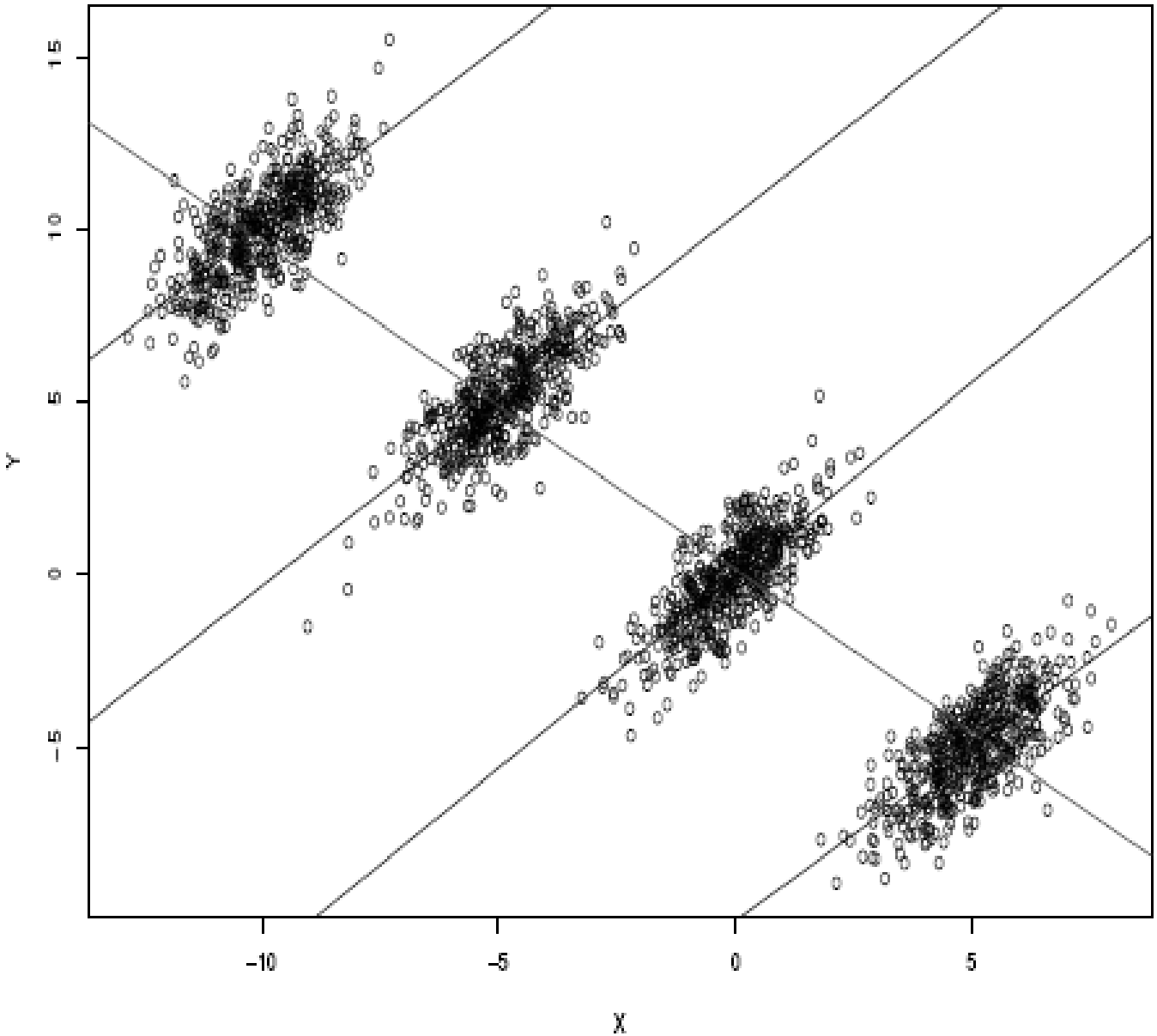
      (4) Unobserved (unmodeled) Coefficient Variability: $\beta_{i\ \&/\mathrm{or}\ t} \neq \beta$

   c) Examples: graphs of heterogeneity bias (specifically: omitted unit-specific factors)

(1) *Inflation bias* (positive β_x, positive correlation of $\overline{x}_i$ & the omitted u_i)



(2) <u>*Sign reversal*</u>: positive β_x, strong negative corr. $\overline{x}_i$ & omitted u_i. (<u>*Attenuation bias*</u> if more moderately negative correlation proportionately.)



(3) If no corr. $\overline{x}_i$ & u_i, no bias, though still inefficient & s.e.'s likely wrong.

B. 1st Defense: Model It! (besides, likely this substance, not nuisance)

1. If, for example, expect some pattern nonsphericity, this likely because you expect some *systematic*…

a) …variation in effect of x_{it} across i,t

(1) => will look like heteroskedasticity if model effect as a constant, β

(2) => want to model the interactive (or group-wise varying) effect:

(a) $\beta_{it} = \gamma_0 + \gamma_1 z_{it} (+\phi_{it})$ (=> linear-interaction model (plus…))

(b) $\beta_{it} = \gamma_i (+\phi_{it})$, or $\beta_{it} = \gamma_t (+\phi_{it})$, or $\beta_{it} = \gamma_s (+\phi_{it})$ (dummy-interax (+…))

b) …dependence of y_{it} on $y_{i,t-1}$, &/or y_{it} on y_{jt}

(1) => what looks like serial &/or spatial error-correlation if fail model the temporal &/or spatial dynamics in outcome, y

(2) => model the temporal &/or spatial/spatiotemporal dynamics:

(a) Temporal: $y_{it} = \alpha_{\{i,t\}} + \beta_{\{i,t\}} x_{i,t} + \rho_{\{i,t\}} y_{i,t-1} + \varepsilon_{it}$

(b)Spatial (Spatiotemporal):

$$y_{it} = \alpha_{\{i,t\}} + \beta_{\{i,t\}} x_{i,t} \left(+\rho_{\{i,t\}} y_{i,t-1}\right) + \theta_{\{i,t\}} \sum_{j \neq i} w_{ij} y_{jt} + \varepsilon_{it}$$

2. Implications of "Model It!" Strategy; first, call all RHS: $\mathbf{X\beta}$

a)

$$\hat{\boldsymbol{\beta}}_{OLS} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'y} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'}\left(\mathbf{X\beta}+\boldsymbol{\varepsilon}\right)$$

$$= \left(\mathbf{X'X}\right)^{-1}\mathbf{X'X\beta} + \left(\mathbf{X'X}\right)^{-1}\mathbf{X'}\boldsymbol{\varepsilon}$$

$$= \boldsymbol{\beta} + \left(\mathbf{X'X}\right)^{-1}\mathbf{X'}\boldsymbol{\varepsilon} \Rightarrow E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta} \text{ if } E\left(\mathbf{X'}\boldsymbol{\varepsilon}\right) = 0 \text{ (as usual)}$$

b)

$$\mathbf{V}\left(\hat{\boldsymbol{\beta}}_{OLS}\right) = \mathbf{V}\left[\left(\mathbf{X'X}\right)^{-1}\mathbf{X'y}\right] = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'V}\left(\boldsymbol{\varepsilon}\right)\mathbf{X}\left(\mathbf{X'X}\right)^{-1}$$

$$= \left(\mathbf{X'X}\right)^{-1}\mathbf{X'}\sigma^2\mathbf{IX}\left(\mathbf{X'X}\right)^{-1}$$

$$= \sigma^2\left(\mathbf{X'X}\right)^{-1}\mathbf{X'X}\left(\mathbf{X'X}\right)^{-1}$$

$$= \sigma^2\left(\mathbf{X'X}\right)^{-1} \text{ (as usual)}$$

c) OLS=BLUE if model right(*spatial); what if imperfect/incomplete?

3. For example, suppose use just linear-interaction, when linear-interaction w/ error (=random coefficients) (This is in Franzese *PA* 2005)

$$\text{Truth: } y_{it} = x_{it}\left(\gamma_0 + \gamma_1 z_{it} + \phi_{it}\right) + \varepsilon_{it}$$

$$\text{Model: } y_{it} = x_{it}\left(\gamma_0^* + \gamma_1^* z_{it}\right) + \varepsilon_{it}^*$$

$$\Rightarrow \hat{\boldsymbol{\beta}}_{OLS} = \left(\begin{bmatrix}\mathbf{x} & \mathbf{xz}\end{bmatrix}'\begin{bmatrix}\mathbf{x} & \mathbf{xz}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{x} & \mathbf{xz}\end{bmatrix}'\mathbf{y}$$

$$= \left(\begin{bmatrix}\mathbf{x} & \mathbf{xz}\end{bmatrix}'\begin{bmatrix}\mathbf{x} & \mathbf{xz}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{x} & \mathbf{xz}\end{bmatrix}'\left[\gamma_0\mathbf{x} + \gamma_1\mathbf{x}\cdot\mathbf{z} + \boldsymbol{\varphi}\cdot\mathbf{x} + \boldsymbol{\varepsilon}\right]$$

a) $$= \begin{bmatrix}\gamma_0 \\ \gamma_1\end{bmatrix} + \left(\begin{bmatrix}\mathbf{x} & \mathbf{xz}\end{bmatrix}'\begin{bmatrix}\mathbf{x} & \mathbf{xz}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{x} & \mathbf{xz}\end{bmatrix}'\left[\boldsymbol{\varphi}\cdot\mathbf{x} + \boldsymbol{\varepsilon}\right]$$

$$\Rightarrow E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta} \text{ if } E\left(\boldsymbol{\varphi}\cdot\mathbf{x}\right) = \mathbf{0},\ E\left(\mathbf{X}'\boldsymbol{\varepsilon}\right) = 0$$

$$\mathbf{V}\left(\hat{\boldsymbol{\beta}}_{OLS}\right) = \mathbf{V}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}\left(\mathbf{x}\cdot\boldsymbol{\varphi} + \boldsymbol{\varepsilon}\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

b) $$= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\sigma^2\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

c) OLS=unbiased, but inefficient coefficients; wrong s.e.'s

d) However, some easy (and now familiar…) fixes…

4. E.g., unit-specific effects or coefficients, but model only part of that parameter heterogeneity (mis-specification: OVB)

$$\text{Truth: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

$$\text{Model: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*$$

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\right)$$

a)

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\gamma} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

$$= \boldsymbol{\beta} + \mathbf{F}_{\mathbf{ZX}}\boldsymbol{\gamma} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \text{ where } \mathbf{F}_{\mathbf{ZX}} \text{ is OLS } \mathbf{Z} \text{ on } \mathbf{X}$$

$$\mathbf{V}\left(\hat{\boldsymbol{\beta}}_{OLS}\right) = \mathbf{V}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right)$$

$$\text{but } \widehat{\mathbf{V}\left(\hat{\boldsymbol{\beta}}_{OLS}\right)} = \hat{\mathbf{V}}\left(\hat{\mathbf{F}}_{\mathbf{ZX}}\hat{\boldsymbol{\gamma}} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right)$$

b)

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

c) I.e., completely-standard omitted-variable bias (OVB).

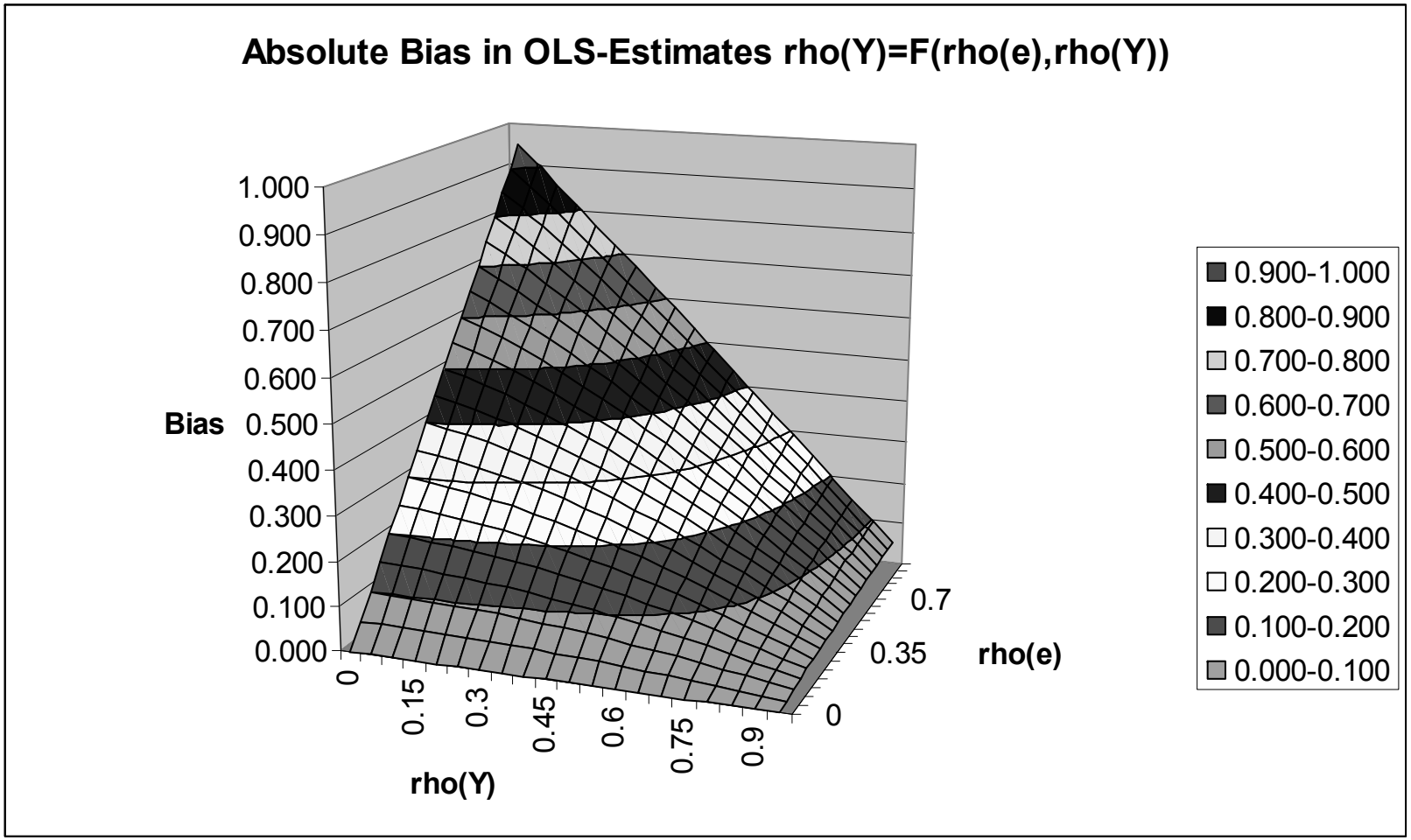
5. (Time-)Serial Dependence:

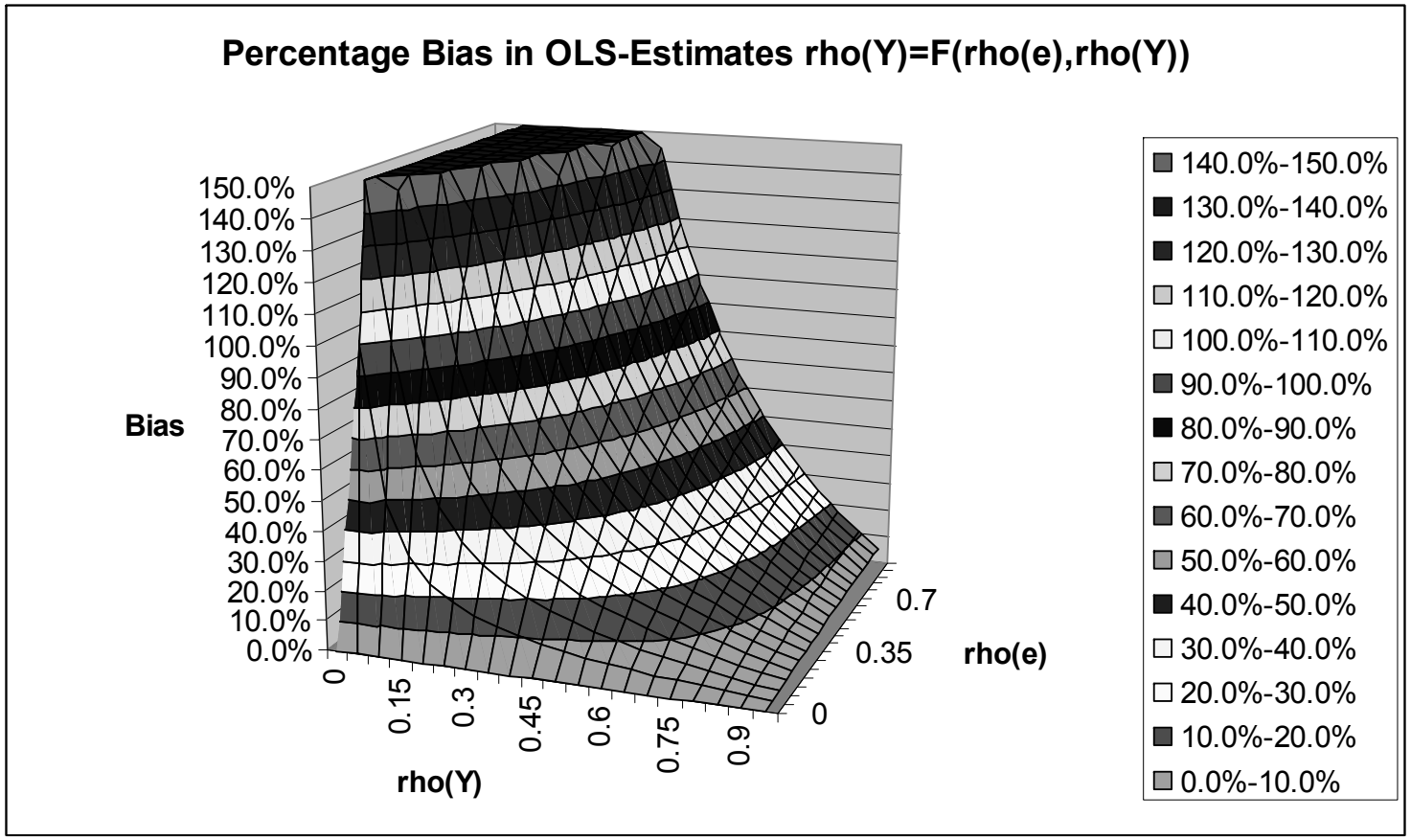
a) If temporal dynamics specified in systematic component sufficiently (no residual/stochastic-component corr. remains, which testable), OLS $\rightarrow^{A}$ BLUE.

b) If insufficient, OLS inconsistent, but still:

(1) Magnitude of the problem: $E\left(\hat{\rho}_y\right) = \rho_y + \rho_\varepsilon\left(1-\rho_y^2\right)\Big/\left(1+\rho_\varepsilon\rho_y\right)$

(2) And can (partially) address s.e. part of problem (as we have seen…)





c) Possible to model temporal dependence in both y & ε by NLS:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \rho_y \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t; \quad \boldsymbol{\varepsilon}_t = \rho_\varepsilon \boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\upsilon}_t; \quad \boldsymbol{\upsilon}_t \sim \left(\mathbf{0}, \sigma_\upsilon^2 \mathbf{I}\right)$$

$$\Rightarrow \mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \rho_y \mathbf{y}_{t-1} + \rho_\varepsilon \boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\upsilon}_t$$

$$= \mathbf{X}\boldsymbol{\beta} + \rho_y \mathbf{y}_{t-1} + \rho_\varepsilon \left(\mathbf{y}_{t-1} - \mathbf{X}_{t-1}\boldsymbol{\beta} - \rho_y \mathbf{y}_{t-2}\right) + \boldsymbol{\upsilon}$$

$$= \left(\mathbf{X}_t - \rho_\varepsilon \mathbf{X}_{t-1}\right)\boldsymbol{\beta} + \rho_y \mathbf{y}_{t-1} + \rho_\varepsilon \left(\mathbf{y}_{t-1} - \rho_y \mathbf{y}_{t-2}\right) + \boldsymbol{\upsilon}$$

d) Note, however, indeterminacy in total=systematic+stochastic, so any two of possible lag y, lag x, lag e => third, etc.

6. Spatial Dependence:

a) Situation more complic. OLS inconsistent even if model spatial-dep fully.

b) However, still generally better to model than to omit it, & we'll talk about redressing the simultaneity in this case later (if we have time).

7. Summary:

a) If can model thry/subst reason for deviation from C(N)LRM, in TSCS data or elsewhere, do so, &, if/insofar as successful, strategy optimal in all regards.

b) Insofar as possible, "Model It!" in model of first-moment, $E(\mathbf{y})$, i.e., systematic component; for two reasons:

(1) Usually, the theory/substance in question regards <u>*systematic*</u> component

(2) Observationally, only info we have on stochastic component (i.e., second moment) conditional on info in & model of first moment (systematic component)

(3) May not be possible or theoretically/substantively correct; &/or could have theory-substance info about second moment (variance). E.g.:

(a) DepVar=average varying # lower-level outcomes => $V(\varepsilon)\sim 1/\#$.

(b) Thry/Subst e.g.: education (or information) not affect response; rather reliability or accuracy or theoretical-explicability of response=> $V(\varepsilon)=f(\text{edu})$

(c) Still "Model It!" (now in 2nd moment, so, i.e., model reduced parameterization of $\mathbf{\Omega}$).

(4) As seen, insofar as fail model fully deviations CLRM, prob's arise essentially as OVB in worst cases, but as "just" inefficiency & wrong s.e.'s else, so…

C. Redresses, mostly partial &/or imperfect, of deficiencies in implementation of the *Model It!*™ strategy

1. "Robust" or "Sandwich" Variance-Covariance Estimators

a) Key Insight:

(1) $\mathbf{\Omega}$ has ½n(n+1)>n elements/parameters; however, for consistent v-cov est:

(2) need consistent est only of $\mathbf{X'\Omega X}$, which only ½k(k+1) elements.

b) Proper $\mathbf{V}\left(\hat{\boldsymbol{\beta}}_{LS}\right)=\sigma^2(\mathbf{X'X})^{-1}\mathbf{X'\Omega X}(\mathbf{X'X})^{-1}$ differs from OLS $\mathbf{V}\left(\hat{\boldsymbol{\beta}}_{OLS}\right)=\sigma^2(\mathbf{X'X})^{-1}$, only insofar as $\mathbf{X'\Omega X}$ differs from $\mathbf{X'X}$, which is only insofar as elements of $\mathbf{\Omega}$ covary with elements of $\mathbf{X'X}$, i.e. ω_{ij} w/ x's, x^2's, &/or cross-products of x's.

c) Visualizing the matrix multiplication, we see that v-cov estimates using LS formula are off by factor of: $\sum_{i,j,s,t} e_{it}e_{js}\left(\mathbf{x}_{it}\mathbf{x}'_{js}\right)-\sum_{i,t} e_{it}^2\mathbf{I}_k$.

d) ∴, we can *fix* our v-cov estimates, i.e. render them *robust*, i.e., <u>consistent</u>, to presence of certain pattern of nonsphericity by replacing $\mathbf{X'\Omega X}$ in formula w/ some $\sum_{i,j,s,t} e_{it}e_{js}\left(\mathbf{x}_{it}\mathbf{x}'_{js}\right)$ configured to reflect that nonsphericity pattern.

$$\hat{\mathbf{V}}_s\left(\hat{\boldsymbol{\beta}}\right)=(\mathbf{X'X})^{-1}\widehat{\mathbf{X'\Omega X}}(\mathbf{X'X})^{-1}\equiv(\mathbf{X'X})^{-1}\hat{\mathbf{Q}}(\mathbf{X'X})^{-1}$$

2. Cases:

a) Pure Heteroskedasiticity (White's): $\hat{\mathbf{Q}} = \sum_i \sum_t e_{it}^2 \left(\mathbf{x}_{it} \mathbf{x}_{it}' \right)$

b) Pure Het. & (Time) Auto-Correlation (HAC) (Newey-West):

$$\hat{\mathbf{Q}} = \sum_i \left(\sum_t e_{it}^2 \left(\mathbf{x}_{it} \mathbf{x}_{it}' \right) \right) + \sum_i \left(\frac{1}{T} \sum_{l=1}^{L} \sum_{t=l+1}^{T} w_l e_t e_{t-l} (\mathbf{x}_t \mathbf{x}_{t-l}' + \mathbf{x}_{t-l} \mathbf{x}_t') \right)$$

where L=max lag-length considered appreciable & $w_t = 1 - \frac{l}{L+1}$

c) Panel Heteroskedasiticity: ?, but White's pure-het more general, so…

d) Panel Het. & (Time) Auto-Correlation: ?, but Newey-West more gen., so…

e) Panel Het. & Contemporaneous (Spatial) Corr. (Beck-Katz PCSE):

$$\hat{\mathbf{Q}} = \mathbf{X}' \left(\frac{\mathbf{E}'\mathbf{E}}{T} \otimes \mathbf{I}_T \right) \mathbf{X}$$

where $\mathbf{E}$=the $T \times N$ matrix estimated residuals

f) Many others possible. Several "cluster" types, e.g., designed for various multilevel/hierarchical data structures (i.e., a panel 'random-effect' structure):

$$\hat{V}\left(\hat{\boldsymbol{\beta}}\right)=\frac{1}{N-k}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{j=1}^{n_c}\left\{\left[\sum_{i=1}^{n_j}e_i\mathbf{x}_i\right]\left[\sum_{i=1}^{n_j}e_i\mathbf{x}_i\right]'\right\}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

where n_j=# obs. i in macro-level (cluster) j, & n_c=# clusters

g) Note: excepting Newey-West, asymptotics for these tend to be in *N* or in some function of *N* and *T*, not in *T*. Many may not work well in TSCS.

h) Some not assuredly "well-behaved" in estimation, so various *kludges*.

i) Small-sample adjusts been suggested for each; may be key in TSCS.

(1) E.g., "[For White's,] Davidson and MacKinnon (1993: 554) strongly suggest a finite-sample correction of replacing e_i^2 by $e_i^2/(1-\mathbf{x}_i(\mathbf{X'X})^{-1}\mathbf{x}_i)$, which scales estimated squared residuals by their variance, or of multiplying by $N/(N-k)$, which inflates estimates by a factor reflecting the number of regressors as a percentage of degrees of freedom. Accumulating simulation work favors their suggestion."

(2) E.g., "As with [White's], a finite-sample (degrees-of-freedom) correction, $[n_c/(n_c-1)][(N-1)/(N-k)]$, is suggested. This inflates standard errors as there, but now multiplicatively further, by a declining function of J. Again, simulations strongly support using such adjustments.

(3) Not clear to me (from `help vcetype`) if and which of these adjustments Stata applies as defaults or options (need manuals).

(4) From many MC's I've seen on Cluster, PCSE, etc., more attention to these small-sample adjustments would be a good thing in TSCS esp.

3. FGLS: Feasible Generalized-Least-Squares

a) Consistent V-Cov Ests only address "inconsistency" of s.e.'s, do not address bias or efficiency of coeff estimates (although require consistent coefficient-estimates for formal properties) or "unbiasedness" and "efficiency" of s.e.'s.

b) To improve efficiency coeff (& s.e.) estimates—still not directly or formally redress any bias concerns arising from other problems, OVB e.g., and still reliant on "first-stage" consistency—we can parameterize and estimate $\hat{\mathbf{\Omega}}$, use it to transform data to such that C(N)LRM applies.

c) Example: *Parks-Kmenta FGLS* for TSCS:

(1) Panel-specific AR(1) in residuals=> *N* parameters

(2) Panel-specific σ_i^2 => *N* parameters

(3) Dyad-specific σ_{ij} => *N(N-1)* parameters (n.b., symmetric)

(4) =>N(N+1) pars => inadvisable unless T>>2N (Beck-Katz '95)

(5) NOTE: Could offer more theoretically structured (& thereby parametrically reduced) structure non-sphericity pattern => greater efficiency & better small-sample properties. E.g., just contemp. corr. => N(N-1) parameters needs T>>N.

d) FGLS properties: consistent & asymptotically efficient.

FGLS: given consistent est $\hat{\mathbf{\Omega}}$, let $\mathbf{P} \equiv \hat{\mathbf{\Omega}}^{-\frac{1}{2}}$, then:

$$\mathbf{Py} = \mathbf{PX\beta} + \mathbf{P\varepsilon} \Rightarrow$$

$$\hat{\boldsymbol{\beta}}_{FGLS} = \left[(\mathbf{PX})' (\mathbf{PX}) \right]^{-1} (\mathbf{PX})' \mathbf{Py}$$

$$\hat{\boldsymbol{\beta}}_{FGLS} = \left[\mathbf{X}'\mathbf{P}'\mathbf{PX} \right]^{-1} \mathbf{XP}'\mathbf{Py} = \left[\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X} \right]^{-1} \mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}$$

$\Rightarrow$ consistent, asympt'ly efficient if $\mathrm{C}(\mathbf{X}, \boldsymbol{\varepsilon}) = 0$

$$\widehat{V\left(\hat{\boldsymbol{\beta}}_{FGLS}\right)}_{FGLS} = \left[\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X} \right]^{-1} \mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{V}(\mathbf{y})\hat{\mathbf{\Omega}}^{-1}\mathbf{X} \left[\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X} \right]^{-1}$$

$$\widehat{V\left(\hat{\boldsymbol{\beta}}_{FGLS}\right)}_{FGLS} = \sigma^2 \left[\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X} \right]^{-1} \mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\hat{\mathbf{\Omega}}\hat{\mathbf{\Omega}}^{-1}\mathbf{X} \left[\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X} \right]^{-1}$$

$$\widehat{V\left(\hat{\boldsymbol{\beta}}_{FGLS}\right)}_{FGLS} = \sigma^2 \left[\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X} \right]^{-1} \mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X} \left[\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X} \right]^{-1}$$

$$\widehat{V\left(\hat{\boldsymbol{\beta}}_{FGLS}\right)}_{FGLS} = \sigma^2 \left[\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X} \right]^{-1}$$

$\Rightarrow$ "consistent and asympt'ly efficient" (as above)

Heterogeneity in TSCS

IV. Notation, an almost most-general (linear) model:

$$y_{it} = \alpha_{it} + \mathbf{x}'_{it}\boldsymbol{\beta}_{it} + \varepsilon_{it};\ \ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma});\ \ i = 1..N,\ t = 1..T,\ n = NT$$

A. Summary:

1. If Gauss-Markov applies, i.e. if C(N)LRM, => OLS=BLUE.
2. If G(N)LRM, GLS=BLUE & FGLS=asymptotically BLUE.
3. TSCS=collection of time-series, so all may know regarding TS models applies (w/ approp care to respect breaks b/w units; e.g., $y_{2,1}$(t-1)$\neq y_{1,T}$).
4. Departure from C(N)LRM lies in plausibility key assumptions:

   a) Parameter Stability & Spherical Errors (homoskedasticity+uncorrelated):

   b) => First Line of Defense, Always & Everywhere: ***Model It!***™ (besides, likely this substance, not nuisance)

   c) If model not right, or not enough, some decent properties may still hold. From theory-evaluation perspective, worry about unmodeled heterogeneity only insofar as failure to model adequately biases or otherwise worsens estimates of what can model/understand or certainty-estimates thereof.

B. Implications of *Model It!*™ Strategy (read as a flow-chart left-to-right)

Model It!™ Adequacy	Second-Moment & Inadequacy Variance-Covariance Structure		Implications for OLS Properties
Model E(**y**) Sufficient	V(**e**) Spherical		OLS is BLUE
	V(**e**) Nonspherical	Nonsphericity Unrelated **xx'**	OLS **b** unbiased, inefficient; OLS V(**b**) unbiased, inefficient
		Nonsphericity Related **xx'**	OLS **b** unbiased, inefficient; OLS V(**b**) biased, inefficient
Model E(**y**) Insufficient	Unmodeled **b** het unrelated **X**	V(**e**) Will be Nonspherical & Related **xx'**	OLS **b** unbiased, inefficient; OLS V(**b**) biased, inefficient
	Unmodeled **b** het related **X**	V(**e**) Will be Nonspherical & Related **xx'**	OLS **b** biased, inefficient; OLS V(**b**) biased, inefficient

[INTERRUPTION: SOME TESTS FOR VARIOUS FORMS NON-SPHERICITY]

White's General Test:

c. White test if form of Heteroskedasticity is unknown:

H0: $V\left[\varepsilon_i \mid x_i\right] = \sigma^2$

Ha: $V\left[\varepsilon_i \mid x_i\right] = \sigma_i^2$

Estimate the model under H0
Compute squared residuals: e_i^2

Use squared residuals as dependent variable of auxiliary regression: RHS: all regressors, their quadratic forms and interaction terms:

$$e_i^2 = \delta_0 + \delta_1 x_{i2} + \ldots + \delta_{k-1} x_{ik} + \delta_{k-1} x_{i2}^2 + \delta_{k+1} x_{i2} x_{i3} + \ldots + \delta_q x_{ik}^2 + \xi_i$$

Compute White statistic from R² of auxiliary regression: $n * R^2 \xrightarrow{a} \chi^2_{(q)}$

Use one-sided test and check if n*R² is larger than 95% quantile of Chi²

Many, Many Heteroskedasticity-Test Strategies:

<u>Szroeter's</u> class of tests: $S = \dfrac{\sum_{i=1}^{NT} h(x_i) e_i^2}{\sum_{i=1}^{NT} e_i^2}$ where $h(x_i)$ is some weight increasing in x_i

King (1982) [not Gary] suggests h(x) = rank(x).

`htest` (downloadable enhanced `hettest`) `szroeter` gives a Q-statistic transformation of *S* such that, under homoscedasticity, *Q* is approximately N(0,1) distributed (Judge 1985:452).

<u>Goldfeld-Quandt</u>: sort e_i by some var. Take high & low sets, $\mathbf{e}_1$ & $\mathbf{e}_2$, (some evidence more power discard middle set), & stat: $\dfrac{\mathbf{e}_1'\mathbf{e}_1 / (n_1 - k)}{\mathbf{e}_2'\mathbf{e}_2 / (n_2 - k)} = F \sim F_{n_1-k, n_2-k}$

<u>Glesjer's</u>: Wald or $NT \times R^2$ set of coefficients in: $\ln e_i^2 = \delta_0 + \delta_1 z_{1i} + \ldots + \delta_{kz} z_{ki} + \upsilon_i$

a. Breusch-Pagan LM test for known form of Heteroskedasticity: groupwise

$$LM = \frac{T}{2}\sum_{i=1}^{n}\left(\frac{s_i^2}{s^2}-1\right)^2$$

s_i^2 =sum of group-specific squared residuals

s^2 = OLS residuals

H0: homoskedasticity ~ Chi² with n-1 degrees of freedom
LM-test assumes normality of residuals, not appropriate if assumption not

b. Likelihood Ratio Statistic

Residuals are computed using MLE (e.g. iterated FGLS, OLS loss of power)

$$-2\ln(\lambda) = (NT)\ln(\sigma^2) - \Sigma\left(T\ln\left(\sigma_i^2\right)\right) \sim \chi^2\left(dF = n-1\right)$$

LM ARCH-Tests (test-statistic $T\times R^2 \sim \chi^2_{(S)}$. For example:

$$e_t^2 = \delta_0 + \delta_1 x_{1_t} + \ldots + \delta_k x_{k_t} + \sum_{s=1}^{S}\lambda_s e_{t-s}^2$$

Tests Spatial Patterns Correlation *(ELAB someday, if time…)*

- "Testing" (measuring) Spatial Association:

$$\text{Moran's I: } I = \frac{N}{S}\frac{\boldsymbol{\varepsilon}'\mathbf{W}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}, \text{ where } S = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}, \text{ if row-nrmlzd: } I = \frac{\boldsymbol{\varepsilon}'\mathbf{W}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}$$

  - See also Anselin's LISA…

- LM's appropriate for LS resids (F&H OxfHndbk):

$$LM_\rho = \frac{\hat{\sigma}_\varepsilon^2\left(\hat{\boldsymbol{\varepsilon}}'\mathbf{W}\mathbf{y}/\hat{\sigma}_\varepsilon^2\right)^2}{G+T\hat{\sigma}_\varepsilon^2}, \text{ and } LM_\lambda = \frac{\left(\hat{\boldsymbol{\varepsilon}}'\mathbf{W}\hat{\boldsymbol{\varepsilon}}/\hat{\sigma}_\varepsilon^2\right)^2}{T}$$

  - But these test v. *iid*, & over-reject…

- Robust LM tests appropriate for SAR v. SAE:

$$LM_\rho^* = G^{-1}\hat{\sigma}_\varepsilon^2\left(\hat{\varepsilon}'Wy/\hat{\sigma}_\varepsilon^2 - \hat{\varepsilon}'W\hat{\varepsilon}/\hat{\sigma}_\varepsilon^2\right)^2, \text{ and } LM_\lambda^* = \frac{\left(\hat{\varepsilon}'W\hat{\varepsilon}/\hat{\sigma}_\varepsilon^2 - \left[T\hat{\sigma}_\varepsilon^2\left(G+T\hat{\sigma}_\varepsilon^2\right)^{-1}\right]\hat{\varepsilon}'Wy/\hat{\sigma}_\varepsilon^2\right)^2}{T\left[1-\frac{T\hat{\sigma}_\varepsilon^2}{G+T\hat{\sigma}_\varepsilon^2}\right]}$$

- Robust Joint Test: $LM_{\rho\lambda} = LM_\lambda + LM_\rho^* = LM_\rho + LM_\lambda^*$

Tests Temporal Correlation

Many, many of these also. My favorite…

LM (Time-)Serial-Correlation Tests (test-statistic $T \times R^2 \sim \chi^2_{(S)}$):

$$e_t = \delta_0 + \delta_1 x_{1_t} + \ldots + \delta_k x_{k_t} + \sum_{s=1}^{S} \lambda_s e_{t-s}$$

…for same reasons Glesjer's/White's favorite among hettests:

Uses familiar, intuitive strategy. Constructive. Powerful.
Plus: valid following LDV model (include among x) unlike, e.g., DW; flexible (detects either MA or AR).

C. Redresses, mostly partial &/or imperfect, of non-spherical V(**e**) deficiencies in implementation of the ***Model It!***™ strategy

1. "Robust" or "Sandwich" Variance-Covariance Estimators
2. FGLS: Feasible Generalized-Least-Squares

Troeger's List of Locuses of Model Heterogeneity

1. Different intercepts:
   first difference models, fixed effects model

2. Different coefficients:
   random coefficient model, SUR model, IA effects

3. Time dependent slopes:
   IA effects

4. Different lag structures:
   no textbook solution available

5. Different dynamics:
   no textbook solution available

V. Textbooks emphasize only 1-3 (A-C) below, and 2 (B) only indirectly:

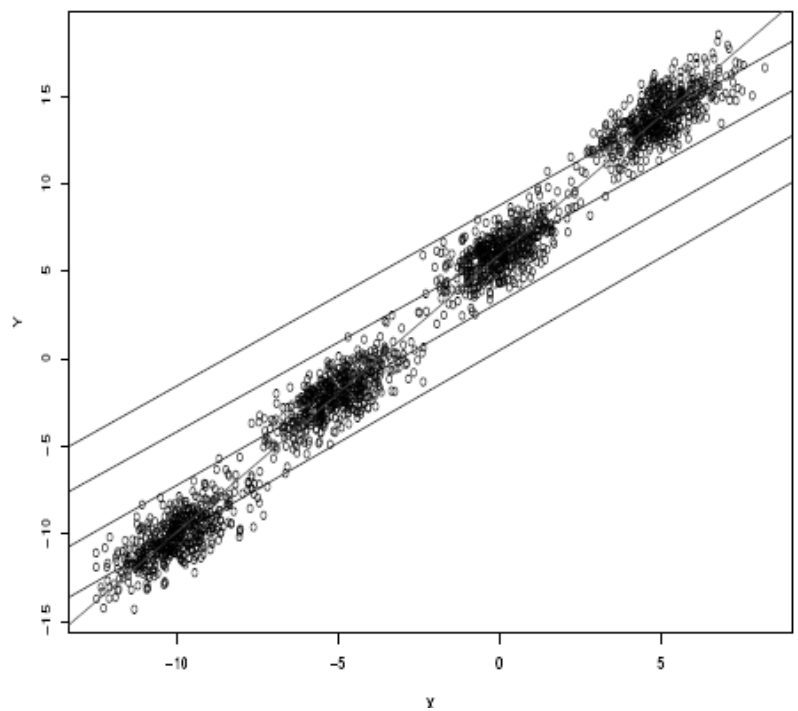
A. Unobserved (unmodeled) Unit (e.g., country) Effects: $\alpha_i \neq \alpha$

B. Unobserved (unmodeled) Time (sub-unit) Effects: $\alpha_t \neq \alpha$

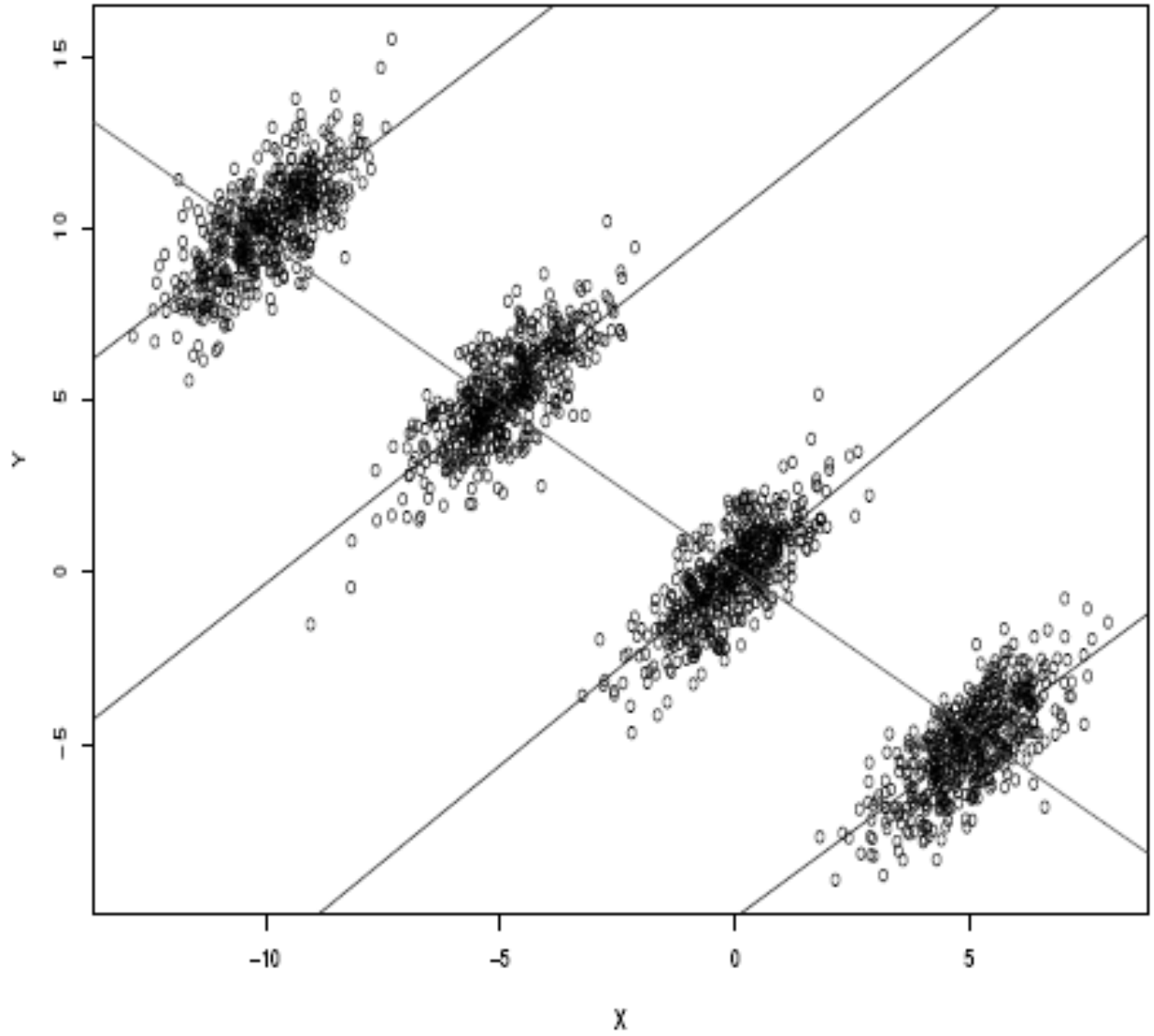
C. Unobserved (unmodeled) Coefficient Variability: $\beta_{i\ \&/\text{or}\ t} \neq \beta$

VI. Examples: graphs of heterogeneity in α

A. <u>*Inflation bias*</u> (positive β_x, positive corr. $\bar{x}_i$ & omitted u_i)

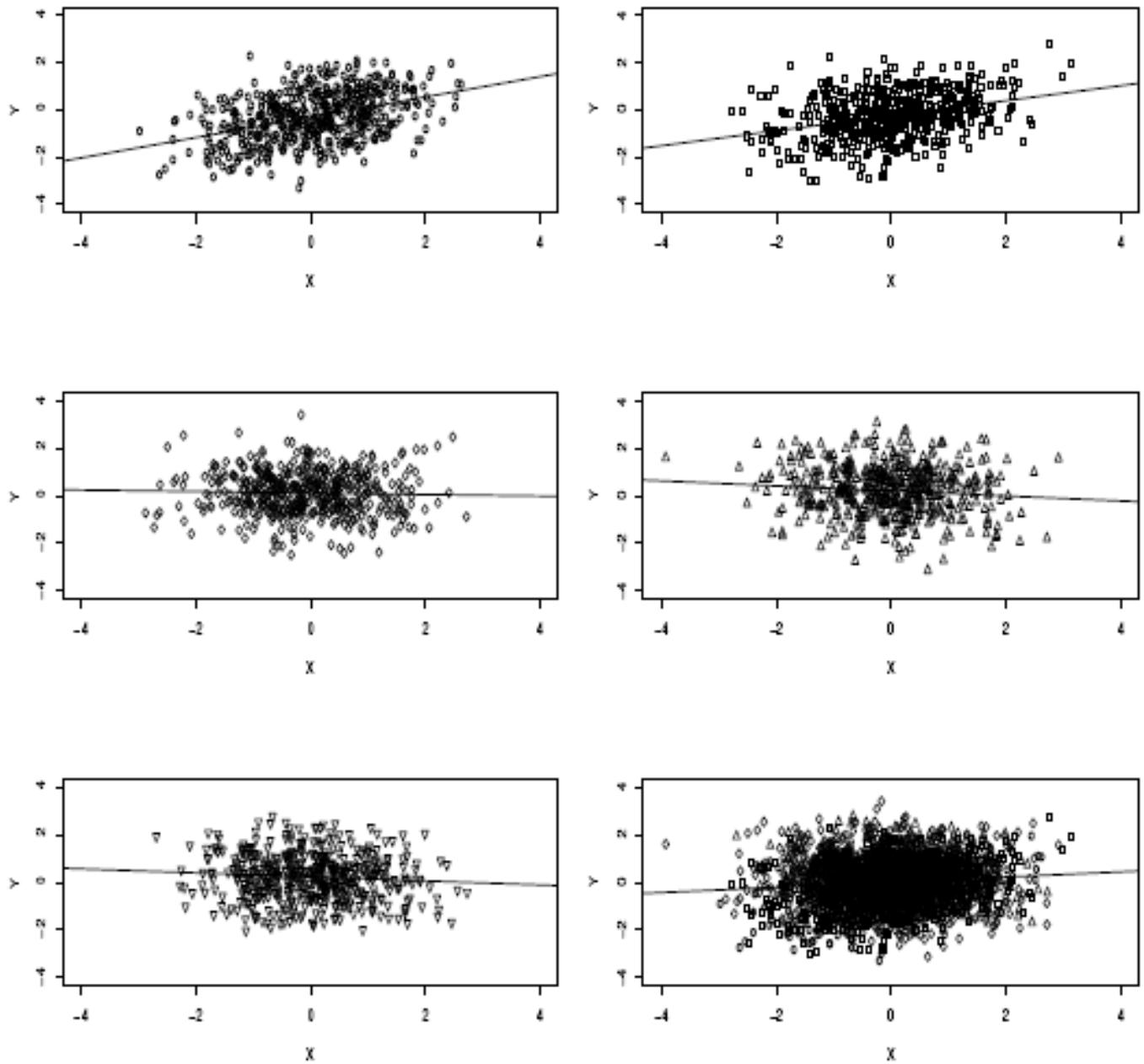


B. <u>*Sign reversal*</u>: positive β_x, strong negative corr. $\bar{x}_i$ & omitted u_i. (<u>*Attenuation bias*</u> if more moderately negative corrproportionately.)



C. If no corr. $\bar{x}_i$ & u_i, no bias, though still ineff & s.e.'s likely wrong.

D. Examples: graphs of heterogeneity in β_i



E. Examples: Heterogeneous Dynamics

1. Unit-Specific AR(IMA) models… May be very important account potentially heterogeneous dynamics:

2. Esp. for slow-moving &/or rarely-changing independent-variables, good estimates depend critically on specifying correct lag structure.

3. Well-known in TS analysis, but, since determining & estimating unit-specific dynamics onerous, most researchers either do not lag IV's or choose arbitrary, uniform lags (mostly one-period). Can have big conseq's:

4. P/T/M (from "Much Left to Do") illustrate using significance of LEFT:

   a) Shifts in value LEFT are frequent but persistent.

5. If institutions matter, may delay policy reactions with less govt autonomy.

6. No generally accepted indicator of opt. lag-length; cands: *t*-, R^2, AIC, BIC

7. P/T/M use un-weighted composite index of these.

8. Result: 11 countries where a change in government has an immediate effect on government spending; 4 countries, Australia, Austria, Germany and Ireland – one-year lag; 2 countries, Italy and Denmark, a two-year lag; 2 countries, Finland and Netherlands, a three-year lag

	model 3.1 uniform lags	model 3.2 optimized lag
Unemployment	1.2146 (0.883) ***	1.2168 (0.0860) ***
GDP per Capita Growth	-0.2604 (0.0403) ***	-0.2608 (0.0401) ***
Dependency Ratio	-0.7384 (0.1267) ***	-0.6655 (0.1229) ***
LEFT (no lags)	0.0002 (0.0041)	
LEFT (optimized lags)		0.0083 (0.0038) **
Christian Democrat Portf.	-0.0418 (0.0093) ***	-0.0338 (0.0088) ***
Trade Openness	0.0507 (0.0290) *	0.0515 (0.0285) *
Low Wage Imports	-0.1488 (0.0436) ***	-0.1567 (0.0420) ***
Foreign Direct Investment	-0.0910 (0.0951)	-0.1217 (0.8604)
N	529	524
R^2	.940	.944
Wald χ^2	13035.32	12310.68
prob> χ^2	0.0000	0.0000
PCSE	yes	Yes
time dummies	no	No
country dummies	yes	Yes

Panel analyses react sensitively to miss-specified lag-structures.

It is theoretically not convincing to assume uniform lags for different units.

Thus:

The precise lag-length should be determined, preferably by theoretical derivation of a hypothesis or empirically by criteria like BIC or AIC

Rejecting a hypothesis because the estimated coefficient turned out to be insignificant while in fact the researcher has wrongly assumed uniform lags would mean blaming the theory for the failures of the methodology.

TROEGER's SUMMARY CONCLUSION

Textbook heterogeneity (FE) is likely to be the least important heterogeneity!

FE "solution" very costly, especially if between-variation is low or theory predicts level effects.

Unit-specific slopes very easy to do. Use fixed coefficients models. (Random coefficient models Beck/Katz 2007 are very a-theoretical.)

Unit-specific dynamics pretty complicated. Perhaps a no go.

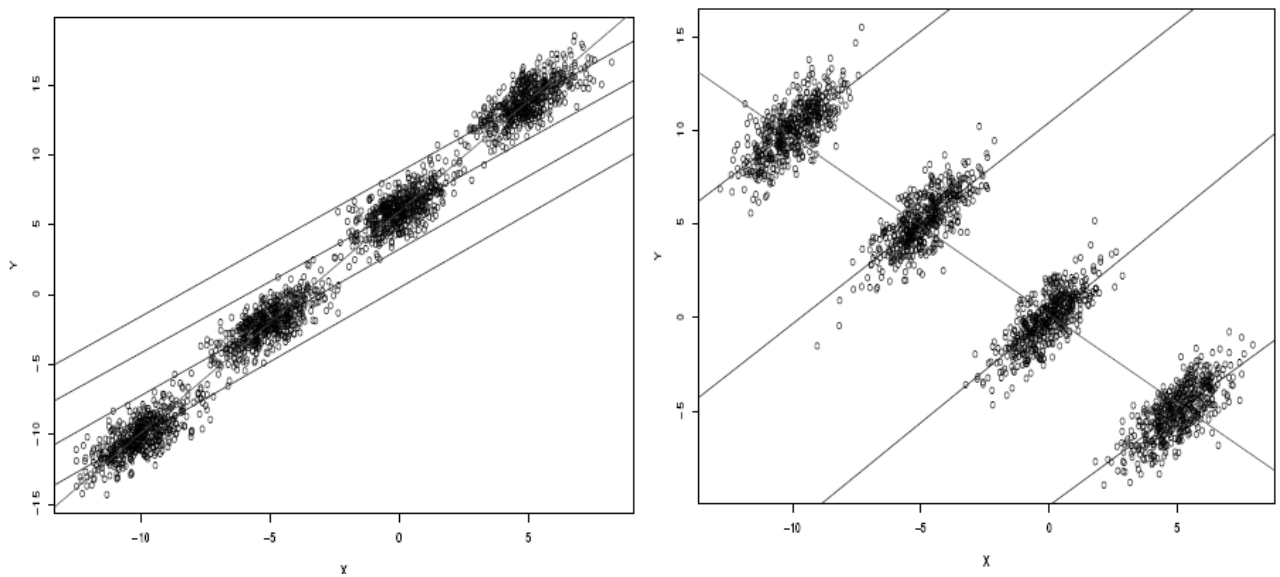
LSDV, a.k.a., FE Models/Estimators in TSCS

VII. (Unmodeled) Unit-Specific "Effects":

A. Unobserved (unmodeled) Unit (e.g., country) Effects: $\alpha_i \neq \alpha$

B. Unobserved (unmodeled) Time (sub-unit) Effects: $\alpha_t \neq \alpha$

VIII. Examples: graphs of heterogeneity in α



A. If Cov($\bar{x}_i$,u_i)=0 no bias, though still ineff & s.e.'s likely wrong.

B. If Cov($\bar{x}_i$,u_i)≠0, then biased, inconsistent, & inefficient.

1.2 Example: Unobserved Country Effects and LSDV

- In the model:

$$Govt\ Spending_{it} = \beta_0 + \beta_1 Openness_{it} + \beta_2 Z_{it} + \varepsilon_{it}$$

  It might be argued that the level of government spending as a percentage of GDP differs for reasons that are specific to each country (e.g., solidaristic values in Sweden). This is also known as cross-sectional heterogeneity.

- If these unit-specific factors are correlated with other variables in the model, we will have an instance of 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 different levels of government spending across countries, but places high demands in terms of data gathering.

- Another way to do this, which may not be as demanding data-wise, is to introduce a set of country dummies into the model.

$$Govt\ Spending_{it} = \alpha_i + \beta_1 Openness_{it} + \beta_2 Z_{it} + \varepsilon_{it}$$

  This is equivalent to introducing a country-specific intercept into the model. Either include a dummy for all the countries but one, and keep the intercept term, or estimate the model with a full set of country dummies and no intercept.

3: *should* try…

4. Equiv, so which…

1.2.1 Time Effects

- There might also be time-specific effects (e.g., government spending went up everywhere in 1973–74 in OECD economies because the first oil shock led to unemployment and increased government unemployment payments). Once again, if the time-specific factors are not accounted for, we could face the problem of bias.

- To account for this, introduce a set of dummies for each time period.

$$Govt\ Spending_{it} = \alpha_i + \delta_t + \beta_1 Openness_{it} + \beta_2 Z_{it} + \varepsilon_{it}$$

- The degrees of freedom for the model are now $NT - k - N - T$. 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*.

- Can also include $(T-1)$ year dummies for time effects. These give the difference between the predicted causal effect from $\mathbf{x}_{it}\boldsymbol{\beta}$ and what you would expect for that year. There has to be one year that provides the baseline prediction.

DLMFE: model?

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.

ΔR^2 test…

1.4.1 How to do this test in Stata?

- After the `regress` command you 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`

1.4.2 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.

Note: Stata's `testparm` useful…

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

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

Note: In this sense, fixed-effect model philosophically less consistent with hyper-population view than random-effects.

- We can rewrite this in vector form as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{e} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \alpha_1^* + \begin{bmatrix} \mathbf{0} \\ \mathbf{e} \\ \vdots \\ \mathbf{0} \end{bmatrix} \alpha_2^* + \ldots + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{e} \end{bmatrix} \alpha_N^* + \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \beta + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_N \end{bmatrix} \quad (3.2)$$

where

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

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

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

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

Note: Hsiao's unfortunate notational choice of **e** for **i**.

Note: could easily combine with panel-heteroskedasticity by $\sigma_{u_i}^2$.

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

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

- 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:

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

- 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.

Note why this therefore also sometimes called “within estimator”

- 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.

Note Troeger's point about limited nature this heterogeneity.

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

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

- 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:

$$\begin{aligned} \mathbf{Q}\mathbf{y}_i &= \mathbf{Q}\mathbf{e}\alpha_i^* + \mathbf{Q}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}\mathbf{u}_i && (3.5) \\ &= \mathbf{Q}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}\mathbf{u}_i && (3.6) \end{aligned}$$

- Running OLS on this regression gives

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

- The variance-covariance matrix is

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

- We can compute an estimate of σ_u^2 as

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

where

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

- Properties of β_{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**.

- A key advantages 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 effects of these variables (perfect collinearity in LSDV).

  ➢ There is an IV approach to produce estimates, but requires some exogeneity assumptions that may not be met in practice.

- The effects of slow-moving variables can also be estimated very imprecisely due to collinearity.

(Yet) Another Way to See How/Why Equivalence Differencing & Dummying

1. Break X' into Two (sets of) variables, X_1 & X_2:

$$Y = X_1 \underset{\sim}{\beta}_1 + X_2 \underset{\sim}{\beta}_2 + e$$

- X_1 is variables (columns) 1-j of X
- X_2 are the remaining (j+1)-k columns (variables)

the $\underset{\sim}{\beta}_1$ & $\underset{\sim}{\beta}_2$ are the vectors of their coefficients.

2. Recall the "normal equations"

$$(X'X)\underset{\sim}{b} = X'\underset{\sim}{Y}$$

, now these can be (are) "broken up" into:

$$\begin{array}{l} (a) \\ (b) \end{array} \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \underset{\sim}{b}_1 \\ \underset{\sim}{b}_2 \end{bmatrix} = \begin{bmatrix} X_1'Y \\ X_2'Y \end{bmatrix}$$

, which, solving by pre-multiplying both sides by $[X'X]^{-1}$ gives

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'\underset{\sim}{Y} \\ X_2'\underset{\sim}{Y} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{b}_1 \\ \underset{\sim}{b}_2 \end{bmatrix}$$

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'\underset{\sim}{Y} \\ X_2'\underset{\sim}{Y} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{b}_1 \\ \underset{\sim}{b}_2 \end{bmatrix}$$

3. By rules of inverting partitioned matrices (chapt. 2)

$$\underset{\sim}{b}_1 = (X_1'X_1)^{-1}X_1'\underset{\sim}{Y} - (X_1'X_1)^{-1}X_1'X_2 b_2$$

$$= A_1\underset{\sim}{Y} - A_1X_2\underset{\sim}{b}_2$$

$$= \boxed{\underset{\sim}{b}_1}$$ — some adjustment which involves coeffs from X_2 on X_1 (A_1X_2) times coeffs. on X_2 in whole regression

(from regression of $\underset{\sim}{Y}$ on X_1 only

subst. this into equation (b) & you get:

$$[X_2'X_1(A_1\underset{\sim}{Y} - A_1X_2\underset{\sim}{b}_2) + X_2'X_2\underset{\sim}{b}_2] = X_2'\underset{\sim}{Y}$$

$$X_2'X_1A_1\underset{\sim}{Y} - X_2'X_1A_1X_2\underset{\sim}{b}_2 + X_2'X_2\underset{\sim}{b}_2 = X_2'\underset{\sim}{Y}$$

(but $X_1A_1 \equiv N_1$)

$$X_2'N_1\underset{\sim}{Y} - X_2'N_1X_2\underset{\sim}{b}_2 + X_2'X_2\underset{\sim}{b}_2 = X_2'\underset{\sim}{Y}$$

(gathering up $\underset{\sim}{b}_2$ terms on LHS & $\underset{\sim}{Y}$ terms on RHS)

$$X_2'(I - N_1)X_2\underset{\sim}{b}_2 = X_2'(I - N_1)\underset{\sim}{Y}$$

$$(X_2'M_1X_2)\underset{\sim}{b}_2 = X_2'M_1\underset{\sim}{Y}$$

$$\underset{\sim}{b}_2 = (X_2'M_1X_2)^{-1}X_2'M_1\underset{\sim}{Y}$$

V. D. "Partial" Regression Coefficients

$$\underset{\sim}{b}_2 = (X_2' M_1 X_2)^{-1} X_2' M_1 \underset{\sim}{Y}$$

now, M_1 is symmetric & idempotent so $M_1' M_1 = M_1 M_1 = M_1$; replace M_1 in above by $M_1' M_1$:

$$\underset{\sim}{b}_2 = (X_2' M_1' M_1 X_2)^{-1} X_2' M_1' M_1 \underset{\sim}{Y}$$

$$(C) \quad \underset{\sim}{b}_2 = [(M_1 X_2)'(M_1 X_2)]^{-1} (M_1 X_2)' M_1 \underset{\sim}{Y}$$

M_1 is "residual-maker" for regression of thing it multiplies on X_1,

and $\underset{\sim}{b}_2$ is coefficient from Regression of $(M_1 \underset{\sim}{Y})$ on $(M_1 X_2)$

(equation (C) is in $(Z'Z)^{-1}Z'Y$ form, right? so it's an A matrix times $\underset{\sim}{Y}$, thus regression coefficients)

but $M_1 \underset{\sim}{Y}$ are the residuals from $\underset{\sim}{Y}$ regressed on X_1 & $M_1 X_2$ are the residuals from (each of) the X_2's regressed on (all of) the X_1's.

- Thus, $\underset{\sim}{b}_2$, the coefficients on X_2 are also the coefficients on X_2^* when regressed on $\underset{\sim}{Y}^*$ where the $*$'s mean that all (linear) relationships with X_1 has been netted out. This is what it means to say "the effect(s) of X_2, controlling for X_1".

Partial Coefficients -- one important example

- suppose we consider X_1 to be just the constant (the vector of ones) and X_2 to be all the other variables. Then, just as before:

$$\underset{\sim}{b}_2 = [(M_1 X_2)'(M_1 X_2)]^{-1}(M_1 X_2)' M_1 \underset{\sim}{Y}$$

- M_1 here is the "residual maker" regressing what it multiplies on the constant only. What do you get regressing any (matrix) vector, Z, on a constant? A coefficient of $\bar{Z}$. Why?

We did this way back in Week 4:

indirectly

$$\underset{a}{\text{Min}}\ (Z - b \cdot \underset{\substack{\uparrow \\ \text{vector} \\ \text{of ones}}}{c})'(Z - b \cdot c) = \underset{a}{\text{Min}} \sum_{i=1}^{n} (Z - b)^2$$

$$\Rightarrow b = \bar{Z}$$

So, then M_1 here nets out the mean of $\underset{\sim}{Y}$ and of all X_2 columns. Thus $\underset{\sim}{b}_1$ in OLS mult. regression including a constant, nets out the means of all X's & Y's.

the coeff on the constant

Troeger summarizes well the "To FE or Not To FE" dilemma:

What it does:

- Eliminates the omitted variables bias of time-invariant unit specific variables

What it does not:

- Does not control away other problems of unit heterogeneity: unobserved time varying variables, slope heterogeneity, unit specific dynamics and lag structures

Problems:

- does not allow estimating the effect of time-invariant variables
- estimator inefficient: likely to obtain estimates that largely deviate from truth, EVEN IF estimates are unbiased
- Does only use within information

Yet,

not controlling for unit effects leads to biased estimates if unit effects exist and are correlated with any of the regressors

Note the Estimator Options So Far:

1. <u>POOLED</u>: $y_{it} = a + \mathbf{x}'\mathbf{b}^p + e_{it}$

2. <u>BETWEEN</u>: $\bar{y}_i = \bar{a}_i + \bar{\mathbf{x}}_i'\mathbf{b}^b + \bar{e}_i$

3. <u>WITHIN (FIXED-EFFECTS (LSDV)</u>: $\left(y_{it} - \bar{y}_i\right) = \left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)' \mathbf{b}^w + \left(e_{it} - \bar{e}_i\right)$

4. <u>UNIT-BY-UNIT</u>: $y_{it} = a_i + \mathbf{x}'\mathbf{b}_i^u + e_{it}$

(Discuss difference between Unit-by-Unit pooled or estimated separately…) A fifth option in this family is to come:

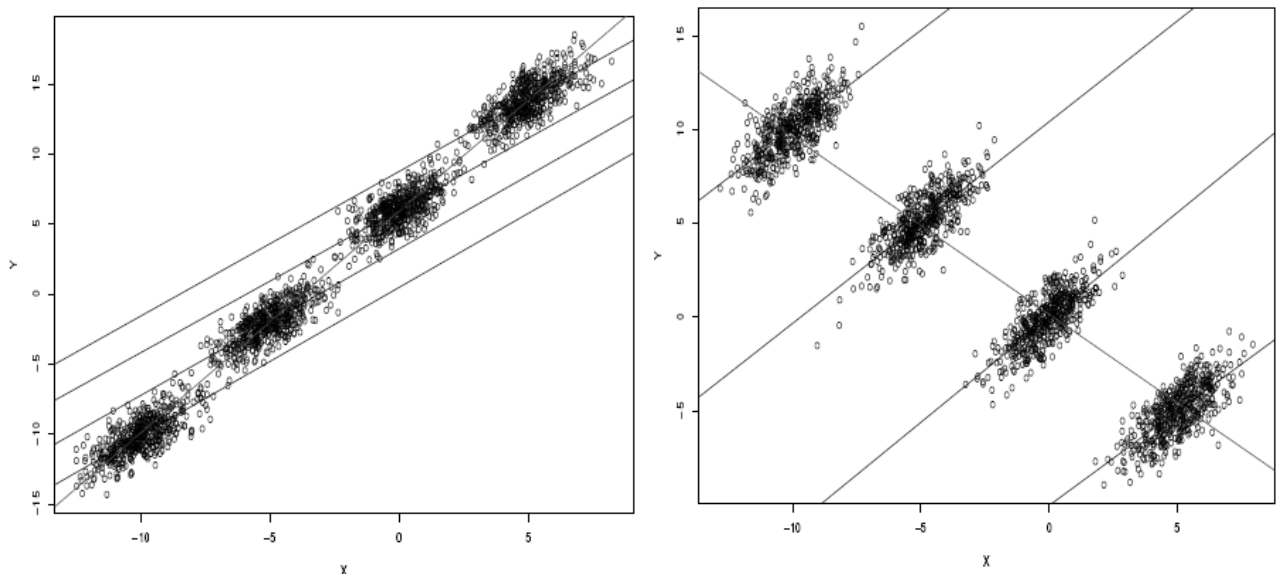
5. <u>RANDOM-EFFECTS</u>

Generally, all can be shown to be weighted averages of each other, each making different compromises. (More on that later if time.)

RE Models/Estimators in TSCS

IX. Issue: Unit (or Time-Period) Heterogeneity: $y_{it} = \alpha_i + \alpha_t + \mathbf{x}\boldsymbol{\beta} + \varepsilon_{it}$

X. Examples: graphs of heterogeneity in α



A. If Cov($\bar{x}_i$,u_i)=0 no bias, though still ineff & s.e.'s likely wrong.

B. If Cov($\bar{x}_i$,u_i)≠0, then biased, inconsistent, & inefficient.

C. Random-effects estimators essentially properly address just 1st.

XI. Substantive/Conceptual issues:

A. Fixed Effects (FE):

1. Appropriate if view unit-specific effects fixed (across repeated samples), estimable amounts, for each cross-section unit. Sweden will always have intercept of 1.2 units, e.g. If somehow took another sample, the intercept for Sweden would be 1.2 again.

2. Costs of FE:

   a) Cannot include ind vars that don't vary over time w/in units (e.g., individuals' demographics, countries' institutions, …).

   b) Highly inefficient: estimates a LOT of parameters. It uses $1/T$ of the degrees of freedom available. Small T, this enormously costly.

   c) Slowly/rarely changing variables highly co-linear w/ *fixed* effects => their effects very imprecisely, unstably, unsurely estimated.

   d) Costs of FE are one reason that people use RE.

   e) Only intercept shift; no other heterogeneity (well) controlled.

   f) Note: many of arguments against hold more for panels than TSCS.

B. Random Effects (RE):

1. If, however, our model estimated in some TSCS of limited N is to generalize to a type of units, not to estimate factual for these units, then these unobserved unit-specific effects should be random.

2. We cannot estimate intercept for each country because we don't have, indeed logically could not have, all of them.

3. May instead wish to estimate coefficients on substantive regressors well, accounting possible country-specific effects that would enter as a random shock from a known distribution: **random effects**.

4. Very efficient relative to FE: need only distributional parameters.

5. Costs:

   a) Must assume a lot: no correlation of random effects with regressors.

   b) Must assume a lot: no correlation among random components.

   c) Addresses only parts of concerns re: unmodeled heterogeneity: the lesser parts in some ways (as per Troeger's comments…)

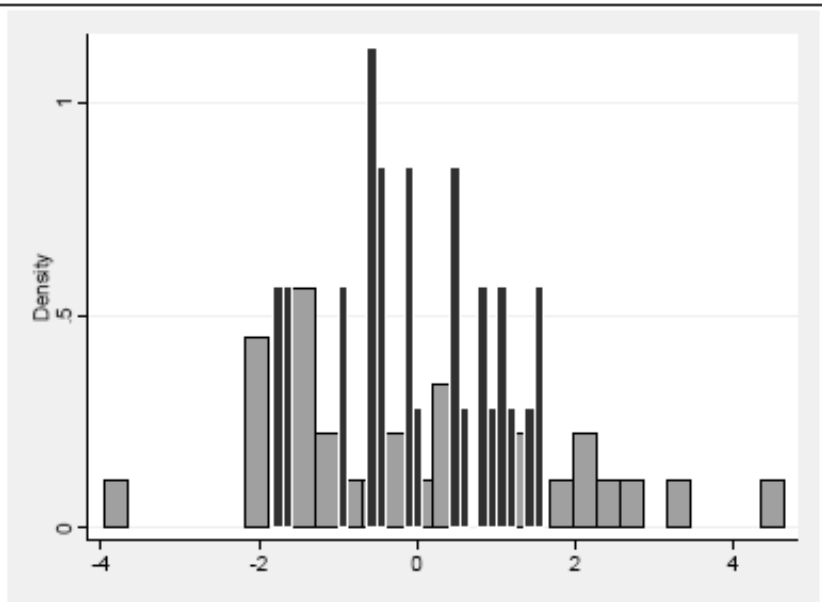
   d) MLE of RE/RC, which best perform, strong Normality assumption.

6. Part of how FE manifests is tendency to "pick up" too much heterogeneity and call it unit-fixed &, in LSDV case, systematic.

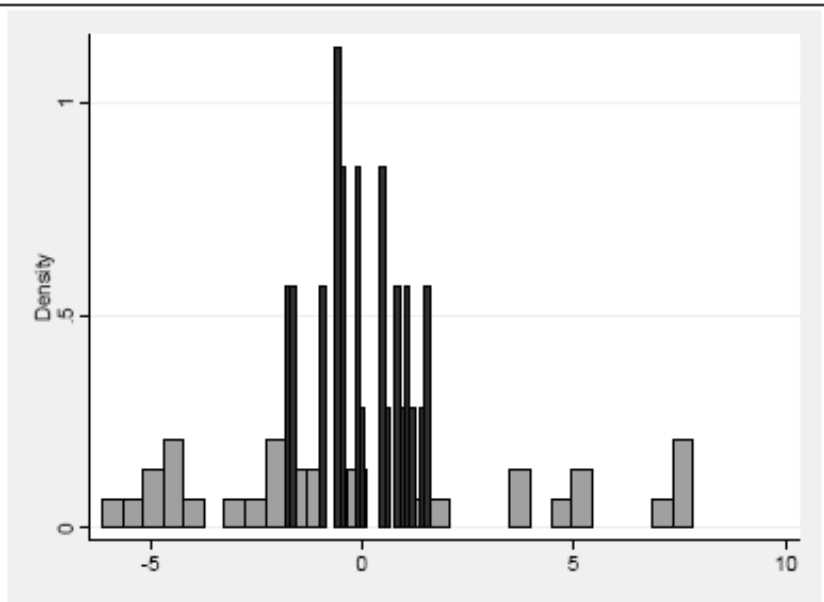
a) I.e., the "sweep" sweeps both fixed & stochastic unit-specific effects.

b) I.e., classic "overfitting"—another way see incidental param problem.

c) Troeger's MC's illustrate the problem: note severe overdispersion of estimated relative to actual unit-specific effects.

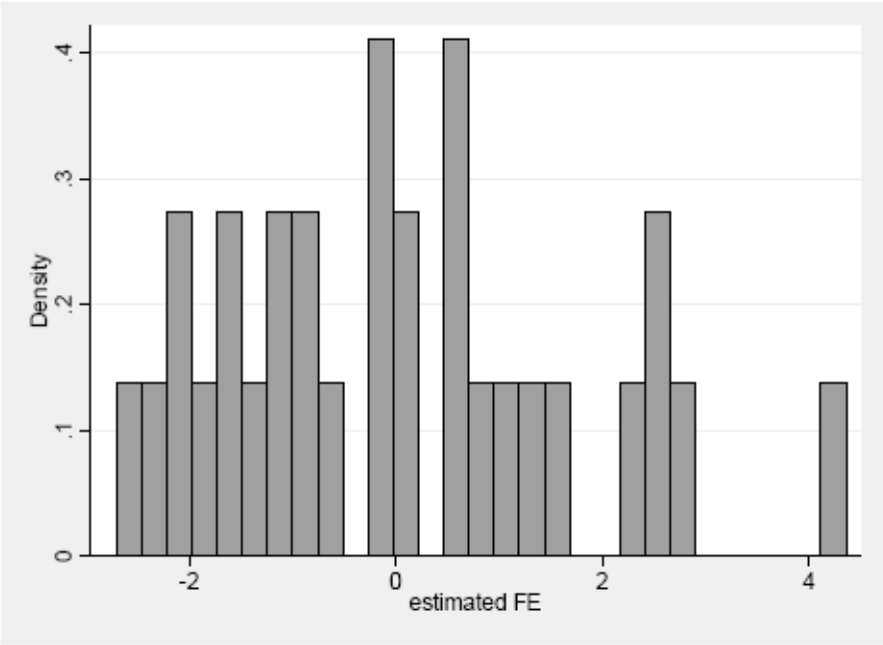


black: true FE, grey: estimated FE
Settings: FE in DGP ~ N(0,1), 1 RHS variable, SD(within)=SD(between)=1

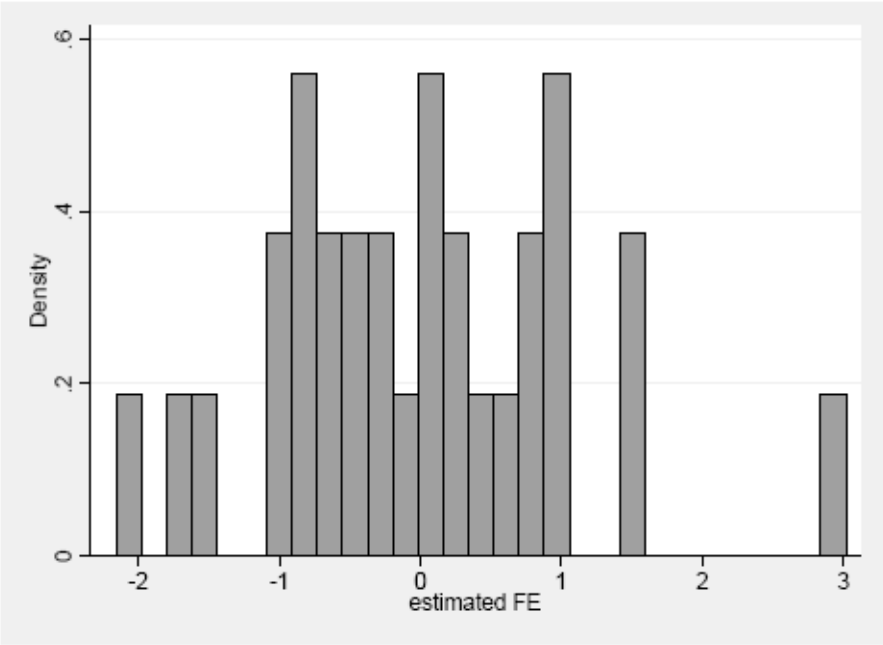


black: true FE, grey: estimated FE
Settings: FE in DGP ~N(0,1), 1 RHS variable SD(within)=1 SD(between)=3

Can be even worse, in fact. Will even find fixed-effects where they ain't:



Settings: no FE in DGP, 1 RHS variable, SD(within)=SD(between)=1



Settings: no FE in DGP, 3 RHS variables, SD(within)=SD(between)=1

Notice: in both this & previous case, not obviously biased, but highly inefficient. In limited (in *T*) samples, can be *not* "mere inefficiency" issue.

XII. A General Random-Effects (Error-Components, HLM) Model:

$$y_{it} = \mathbf{x}\boldsymbol{\beta} + \alpha_i + \lambda_t + u_{it} = \mathbf{x}\boldsymbol{\beta} + v_{it}$$

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

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

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

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

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

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

Mean-zero for each component.

No correlation across components.

Constant variance each component.

Regressors exogenous each component.

Note: second & last lines are crucial; w/o them (or some sufficient replacement), model (becomes) inestimable.

Note: last line implies "only" a GLRM departure from CLRM.

Note: sphericity w/i & across error components implies:

$$\mathrm{var}[y_{it} | \mathrm{x}_{it}] = \sigma_y^2 = \sigma_\alpha^2 + \sigma_\lambda^2 + \sigma_u^2$$

XIII. The Typical RE Model:

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

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

- We can rewrite this in vector form:

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

where

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

I.e., the typical RE model has just two error components: a unit-specific component and the unit-time specific component.

XIV. RE is an example of the G(N)LRM, where:

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

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

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

- Note that

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

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

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

A. So FE estimator would be consistent, but inefficient & wrong V(**b**) est.

B. Since RE Model if G(N)LRM, suggests FGLS estimator:

- The normal equations for the GLS estimator are

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

- We could write the GLS estimator simply as

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

, where

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

and $\psi = \dfrac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$.

1. Recall that Q here is the FE estimator “sweep” matrix.
2. So, RE doesn’t sweep all the unit-averages out, only 1-ψ of them.
   a) Typical GLS intuition for correlated obs: partial-difference the correlation.

C. As usual, then, FGLS applies OLS to transformed **y** & **X**. Here:

$$y_{it}^{*} = y_{it} - \theta \bar{y}_i \text{ , where } \theta \equiv \frac{\sigma_u}{\sqrt{T\sigma_\alpha^2 + \sigma_u^2}}$$

D. Can show that either of these expressions implies that the RE estimator is a weighted average of the pooled & within estimates:

$$\hat{\boldsymbol{\beta}}_{\text{Model}} = \left[\Sigma_{i=1}^{N} \mathbf{X}_i' \left(\mathrm{I} - \frac{\hat{\gamma}_{i,\text{Model}}}{T_i} \mathbf{ii}' \right) \mathbf{X}_i \right]^{-1} \left[\Sigma_{i=1}^{N} \mathbf{X}_i' \left(\mathrm{I} - \frac{\hat{\gamma}_{i,\text{Model}}}{T_i} \mathbf{ii}' \right) \mathbf{y}_i \right]$$

$\hat{\gamma}_{\text{Model}} = 1$ for fixed effects.

$\hat{\gamma}_{i,\text{Model}} = \sqrt{\dfrac{T_i \hat{\sigma}_u^2}{\hat{\sigma}_\varepsilon^2 + T_i \hat{\sigma}_u^2}}$ for random effects.

As $T_i \to \infty, \hat{\gamma}_{i,RE} \to 1$, random effects becomes fixed effects

As $\hat{\sigma}_u^2 \to 0, \hat{\gamma}_{i,RE} \to 0$, random effects becomes OLS (of course)

As $\hat{\sigma}_u^2 \to \infty, \hat{\gamma}_{i,RE} \to 1$, random effects becomes fixed effects

E. In fact, all these various estimators are related thus:

(The following is from Greene, *Econometric Analysis*, ch. 13.)

$$\mathbf{S}_{xx}^{total} = \mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between} \quad \text{and} \quad \mathbf{S}_{xy}^{total} = \mathbf{S}_{xy}^{within} + \mathbf{S}_{xy}^{between}.$$

Therefore, there are three possible least squares estimators of $\boldsymbol{\beta}$ corresponding to the decomposition. The least squares estimator is

$$\mathbf{b}^{total} = \left[\mathbf{S}_{xx}^{total}\right]^{-1}\mathbf{S}_{xy}^{total} = \left[\mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between}\right]^{-1}\left[\mathbf{S}_{xy}^{within} + \mathbf{S}_{xy}^{between}\right]. \tag{13-11}$$

The within-groups estimator is

$$\mathbf{b}^{within} = \left[\mathbf{S}_{xx}^{within}\right]^{-1}\mathbf{S}_{xy}^{within}. \tag{13-12}$$

This is the LSDV estimator computed earlier. [See (13-4).] An alternative estimator would be the between-groups estimator,

$$\mathbf{b}^{between} = \left[\mathbf{S}_{xx}^{between}\right]^{-1}\mathbf{S}_{xy}^{between} \tag{13-13}$$

(sometimes called the **group means estimator**). This least squares estimator of (13-10c)

$$\mathbf{S}_{xy}^{within} = \mathbf{S}_{xx}^{within}\mathbf{b}^{within} \quad \text{and} \quad \mathbf{S}_{xy}^{between} = \mathbf{S}_{xx}^{between}\mathbf{b}^{between}.$$

Inserting these in (13-11), we see that the least squares estimator is a **matrix weighted average** of the within- and between-groups estimators:

$$\mathbf{b}^{total} = \mathbf{F}^{within}\mathbf{b}^{within} + \mathbf{F}^{between}\mathbf{b}^{between}, \tag{13-14}$$

where

$$\mathbf{F}^{within} = \left[\mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between}\right]^{-1}\mathbf{S}_{xx}^{within} = \mathbf{I} - \mathbf{F}^{between}$$

The form of this result resembles the Bayesian estimator in the classical model discussed in Section 16.2. The resemblance is more than passing; it can be shown [see, e.g., Judge (1985)] that

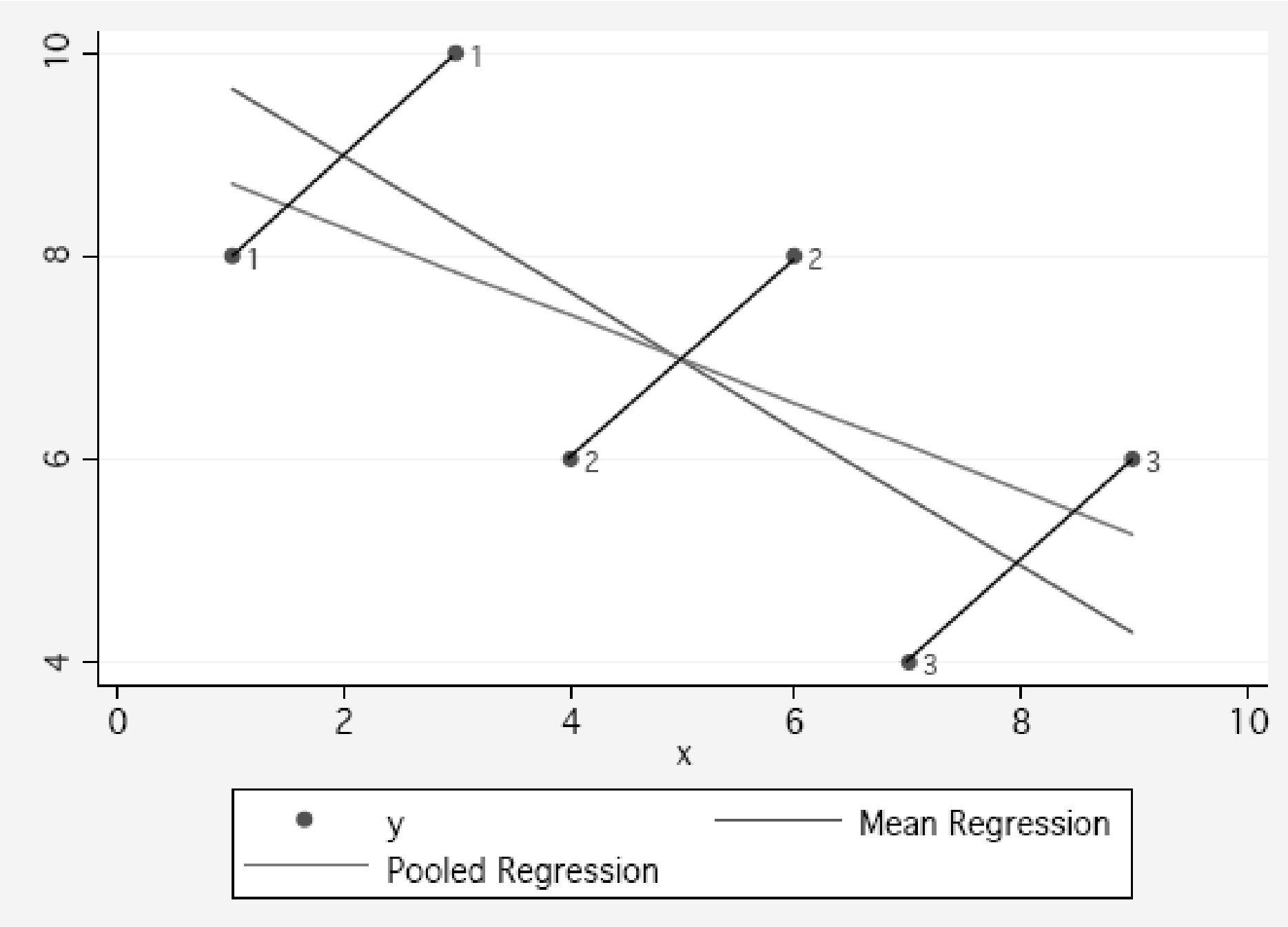
$$\mathbf{F}^{within} = \left\{[\text{Asy. Var}(\mathbf{b}^{within})]^{-1} + [\text{Asy. Var}(\mathbf{b}^{between})]^{-1}\right\}^{-1}[\text{Asy. Var}(\mathbf{b}^{within})]^{-1},$$

which is essentially the same mixing result we have for the Bayesian estimator. In the weighted average, the estimator with the smaller variance receives the greater weight.

--So, e.g., OLS weighs *between* relative to *within* proportionately to shares of total variation at those aggregations.

--That can yield (seeming) too much weight on *between*, as next fig.:

(Figure courtesy of Marco Steenbergen)


10
8
6
4
0
2
4
6
8
10
x
1
1
2
2
3
3
y
Mean Regression
Pooled Regression

It can be shown that the GLS estimator is, like the OLS estimator, a matrix weighted average of the within- and between-units estimators:

$$\hat{\beta} = \hat{\mathbf{F}}^{within}\mathbf{b}^{within} + (\mathbf{I} - \hat{\mathbf{F}}^{within})\mathbf{b}^{between},^{12} \tag{13-23}$$

where now,

$$\hat{\mathbf{F}}^{within} = \left[\mathbf{S}_{xx}^{within} + \lambda\mathbf{S}_{xx}^{between}\right]^{-1}\mathbf{S}_{xx}^{within},$$

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2} = (1-\theta)^2.$$

To the extent that λ differs from one, we see that the inefficiency of least squares will follow from an inefficient weighting of the two estimators. Compared with generalized least squares, ordinary least squares places too much weight on the between-units variation. It includes it all in the variation in $\mathbf{X}$, rather than apportioning some of it to random variation across groups attributable to the variation in u_i across units.

--I like to say *seems* here, because this is assuming RE right model.

--FE, conversely, due to what is essentially overfitting, seems place too much weight on *within* variation (by placing 0 wt on *between*).

To repeat:

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).

So, recall the GLS estimator (in Wawro's notation):

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

, where

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

We didn’t talk about the V-Cov of these GLS estimates yet:

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:

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

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 $\text{var}\left[\sqrt{T}\hat{\boldsymbol{\beta}}_{\text{GLS}}\right] \to \text{var}\left[\sqrt{T}\hat{\boldsymbol{\beta}}_{\text{CV}}\right]$ (assuming our cross-product matrices converge to finite p.d. matrices).

--That is, the first is "smaller" than the second (efficiency), under RE being the right model. And RE→FE as $T\sigma_\alpha^2 \to \infty$.

FGLS: How estimate RE model?

--need estimates of "sweep" terms, which are :

$$y_{it}^{*} = y_{it} - \theta \bar{y}_i \text{ , where } \theta \equiv \frac{\sigma_u}{\sqrt{T\sigma_\alpha^2 + \sigma_u^2}}$$

--to implement 2-stage FGLS, need consistent first-stage.

--since RE model just non-spherical errors, plenty of options for consistent 1st-stages: between (avg), within (FE), pooled (OLS).

-- within estimator (LSDV or unit-mean-differenced) purges α_i by construction, so it's estimated residual-variance is used for $\hat{\sigma}_u^2$.

-- between estimator (group-means regression) contains $\hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2$, so it minus LSDV-based $\hat{\sigma}_u^2$ produces the desired $\hat{\sigma}_\alpha^2$ estimate.

--plug into above, transform, and OLS.

--can get $\hat{\sigma}_\alpha^2 < 0$, but taken to mean no "effects" worry about

- 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:

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

- 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.

Note: multivariate normality, combined with the uncorrelated error-component assumptions amounts to a strong independence assumpt. This what meant earlier by *strong normality* assumptions of ML-RE.

- 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 $N \to \infty$.

MLE generally outperforms FGLS in these kinds of settings; this is very noticeable as the problem grows in complexity, such as in the move from RE to RC… Of course, this relative efficiency of MLE comes from its stronger assumptions & so presumably at price of distributional fragility…

Note the big differences FE vs. RE—

--FE addresses the potential bias issue from unit-specific effects, but inefficient—relatedly, it overfits; in as sense, *because* it overfits, potentially massively so, in fact even infinitely so, relative to some questions (those re: time-invariant regressors, some of which are often of considerable substantive interest: e.g., effects of institutions. Asymptotics best in T.

--RE addresses the efficiency issue, but must assume away the bias issue to get that traction. Asymptotics best in N.

--Can make rather large difference to our estimates and conclusions (even though, as T increases, RE approaches FE). For example:

```
. xtreg sstran L.sstran LL.sstran unem left lftun growthpc depratio
cdem trade lowwage fdi, fe
Fixed-effects (within) regression               Number of obs      =       506
Group variable: cc                              Number of groups   =        17
R-sq:  within  = 0.9756                         Obs per group: min =        19
       between = 0.9952                                        avg =      29.8
       overall = 0.9853                                        max =        31
                                                F(11,478)          =   1737.13
corr(u_i, Xb)  = 0.6209                         Prob > F           =    0.0000
      sstran |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sstran |
         L1. |   1.115156   .0372883    29.91   0.000     1.041887    1.188426
         L2. |    -.18442   .0372824    -4.95   0.000    -.2576776   -.1111624
        unem |   .0356034   .0172374     2.07   0.039     .0017329    .0694738
        left |  -.0005694   .0014259    -0.40   0.690    -.0033712    .0022325
       lftun |   .0002063   .0002349     0.88   0.380    -.0002552    .0006678
    growthpc |  -.1965328   .0107684   -18.25   0.000    -.2176922   -.1753735
    depratio |   .0411629   .0192749     2.14   0.033      .003289    .0790368
        cdem |  -.0023557   .0020135    -1.17   0.243     -.006312    .0016007
       trade |  -.0070589   .0037508    -1.88   0.060    -.0144291    .0003113
     lowwage |   .0005358   .0062747     0.09   0.932    -.0117936    .0128651
         fdi |   .0136313   .0216196     0.63   0.529    -.0288499    .0561124
       _cons |   .5041075   .8362199     0.60   0.547    -1.139014    2.147229
-------------+----------------------------------------------------------------
     sigma_u |   .5312598
     sigma_e |   .5109317
         rho |  .51949773   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(16, 478) =     4.79             Prob > F = 0.0000
. estimates store FEmodel

. xtreg sstran L.sstran LL.sstran unem left lftun growthpc depratio
```

```
cdem trade lowwage fdi, re

Random-effects GLS regression                   Number of obs      =       506
Group variable: cc                              Number of groups   =        17
R-sq:  within  = 0.9738                         Obs per group: min =        19
       between = 0.9990                                        avg =      29.8
       overall = 0.9897                                        max =        31
Random effects u_i ~ Gaussian                   Wald chi2(11)      =  47476.44
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
      sstran |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sstran |
         L1. |   1.212013   .0359035    33.76   0.000     1.141644    1.282383
         L2. |  -.2324231   .0361611    -6.43   0.000    -.3032975   -.1615487
        unem |  -.0265803   .0096796    -2.75   0.006    -.0455519   -.0076086
        left |  -.0005572   .0011681    -0.48   0.633    -.0028466    .0017321
       lftun |   .0002636   .0002115     1.25   0.213    -.0001509     .000678
    growthpc |  -.1780914   .0108472   -16.42   0.000    -.1993515   -.1568313
    depratio |   .0208042   .0107735     1.93   0.053    -.0003114    .0419199
        cdem |     .00122   .0011412     1.07   0.285    -.0010167    .0034566
       trade |   .0035654   .0013232     2.69   0.007      .000972    .0061588
     lowwage |   .0080255   .0040181     2.00   0.046     .0001501    .0159009
         fdi |  -.0308592   .0178517    -1.73   0.084    -.0658479    .0041295
       _cons |   .0557606   .4460443     0.13   0.901    -.8184702    .9299914
-------------+----------------------------------------------------------------
     sigma_u |          0
     sigma_e |   .5109317
         rho |          0   (fraction of variance due to u_i)
. estimates store REmodel
```

Could be much at stake in being able distinguish FE vs. RE…

Hausman test:

--In a (REALLY) large sample, can distinguish pair of estimates where, under null, both are consistent but one more efficient, and, under alternative, the first remains consistent but the second becomes biased.

--Test is basically follows the Wald logic in its structure:

$$Haus = \left(\hat{\boldsymbol{\theta}}_c - \hat{\boldsymbol{\theta}}_e\right)' \left[{}^A\widehat{\text{var}}\left(\hat{\boldsymbol{\theta}}_c\right) - {}^A\widehat{\text{var}}\left(\hat{\boldsymbol{\theta}}_e\right)\right]^{-1} \left(\hat{\boldsymbol{\theta}}_c - \hat{\boldsymbol{\theta}}_e\right) \sim^A \chi^2_k$$

--null in this case is that unit-specific shocks uncorrelated with regressors.

--under that null, both FE & RE are consistent, but RE is efficient.

--under alternative, FE is consistent still, but RE is biased.

Rejection: (supposed to) imply RE rejected in favor of FE….

---Problem 1: can get negative est'd V-Cov diff in samples. Assumed to mean very strong rejection of null.

---*Problem 2*: weak in limited samples.

---**Problem 3**: lots of reasons FE may differ RE (both biased, e.g.).

--Still…wide range uses (e.g., comparing IV's or IV's & OLS) & has that wonderful property: *existence*.

```
. hausman FEmodel REmodel

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
```

```
             |    FEmodel      REmodel       Difference          S.E.
-------------+----------------------------------------------------------------
    L.sstran |    1.115156     1.212013       -.0968571        .0100676
   L2.sstran |     -.18442    -.2324231        .0480031        .0090748
        unem |    .0356034    -.0265803        .0621837         .014263
        left |   -.0005694    -.0005572       -.0000121        .0008179
       lftun |    .0002063     .0002636       -.0000573        .0001022
    growthpc |   -.1965328    -.1780914       -.0184414               .
    depratio |    .0411629     .0208042        .0203586        .0159829
        cdem |   -.0023557       .00122       -.0035756        .0016588
       trade |   -.0070589     .0035654       -.0106244        .0035097
     lowwage |    .0005358     .0080255       -.0074897        .0048194
         fdi |    .0136313    -.0308592        .0444905        .0121952
------------------------------------------------------------------------------
                           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(11) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       27.13
                Prob>chi2 =      0.0044
                (V_b-V_B is not positive definite)
```

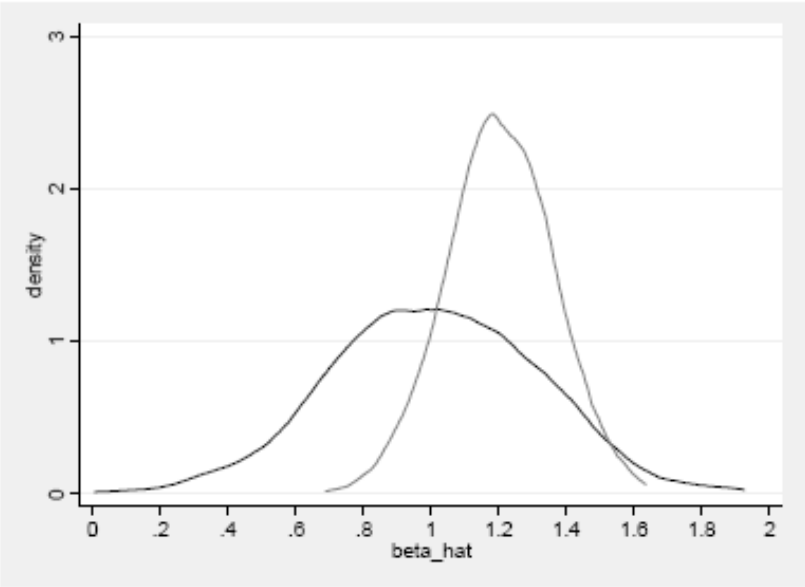
Hausman Test:

Good asymptotic properties and performs well with low frequency of type 1 errors (Baltagi 2001).

Variants of Hausman Test that allow for serial correlation, non-stationarity, heteroskedasticity and other violations of typical assumptions: power analyses & MC studies show these augmentations of Hausman do work better in detecting differences in two estimators under conditions for which they were developed.

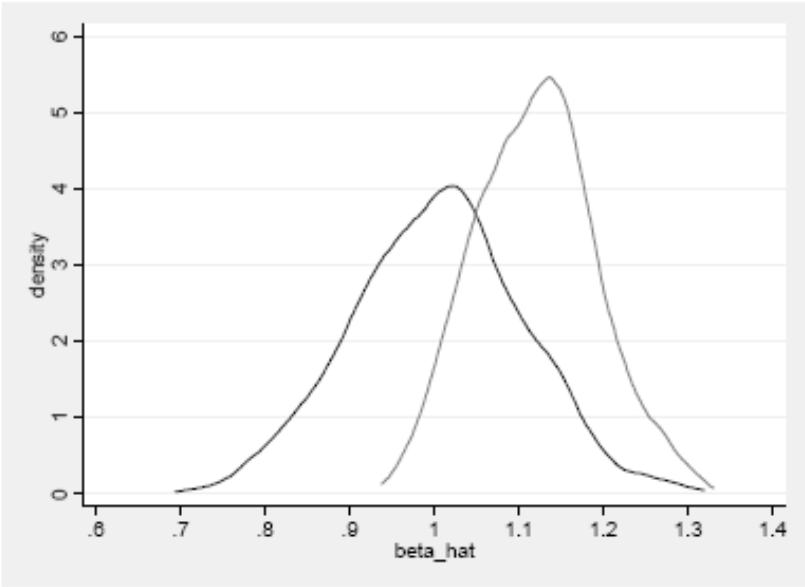
But, in small sample: inefficiency of Hausman test derives in good part from inefficiency of the always-consistent estimator in its tests. (More from Troeger: FE can be badly inefficient …)

Trade-off between efficiency and bias



black: FE; grey: RE
Settings : N=T=10, b/w-ratio=3, DGP u~N(0,1), 1000 repetitions, corr (u,x)=0.4, true beta=1

Trade-off between efficiency and bias



black: FE; grey: RE
Settings : N=T=70, b/w-ratio=3, DGP u~N(0,1), 1000 repetitions, corr (u,x)=0.4, true beta=1

PART 2: Consequences of large sampling variation

Sampling variation of the FE estimates includes values that are 50 to 80 % smaller or larger than the true value

Only 8 percent of the 1000 coefficients (in the extreme tails) turn out to be statistically insignificant.

In an analysis with real world data these deviations from the true value can be quite substantial and create misleading inferences.

The estimated standard errors of the fixed effects estimates are about one third larger than the standard errors of the random effects estimates.

The fixed effects standard errors remain constant for all point estimates regardless of where in the distribution the point estimate lies.

Thus, point estimates that are far away from the true relationship might still be statistically significant.

PART 2: Consequences for the Hausman-test

- Test results are influenced by the trade-off between bias and efficiency
- Hausman test is only powerful in the limit, since FE is consistent (in the limit) – difference of RE and FE estimates only result from biased RE estimates
- In finite samples: differences can result from two sources: biased RE estimates and unreliable FE estimates (because of inefficient estimation)
- Hausman-test mirrors this trade-off: difference of RE and FE estimates / difference in asymptotic variance of RE and FE estimates

→ Test results should be especially unreliable if regressors are both correlated with the unit specific effects and rarely changing

Modeling and Interpreting Interactions

JWAC Mini-Course TSCS

9-11 September 2009

Robert J. Franzese, Jr.

Professor of Political Science,

The University of Michigan, Ann Arbor

1

[The above should be a link to PDF of PowerPoint Presentation on Interaction Terms]

RC (Hierarchical, Multilevel) Models/Estimators

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.

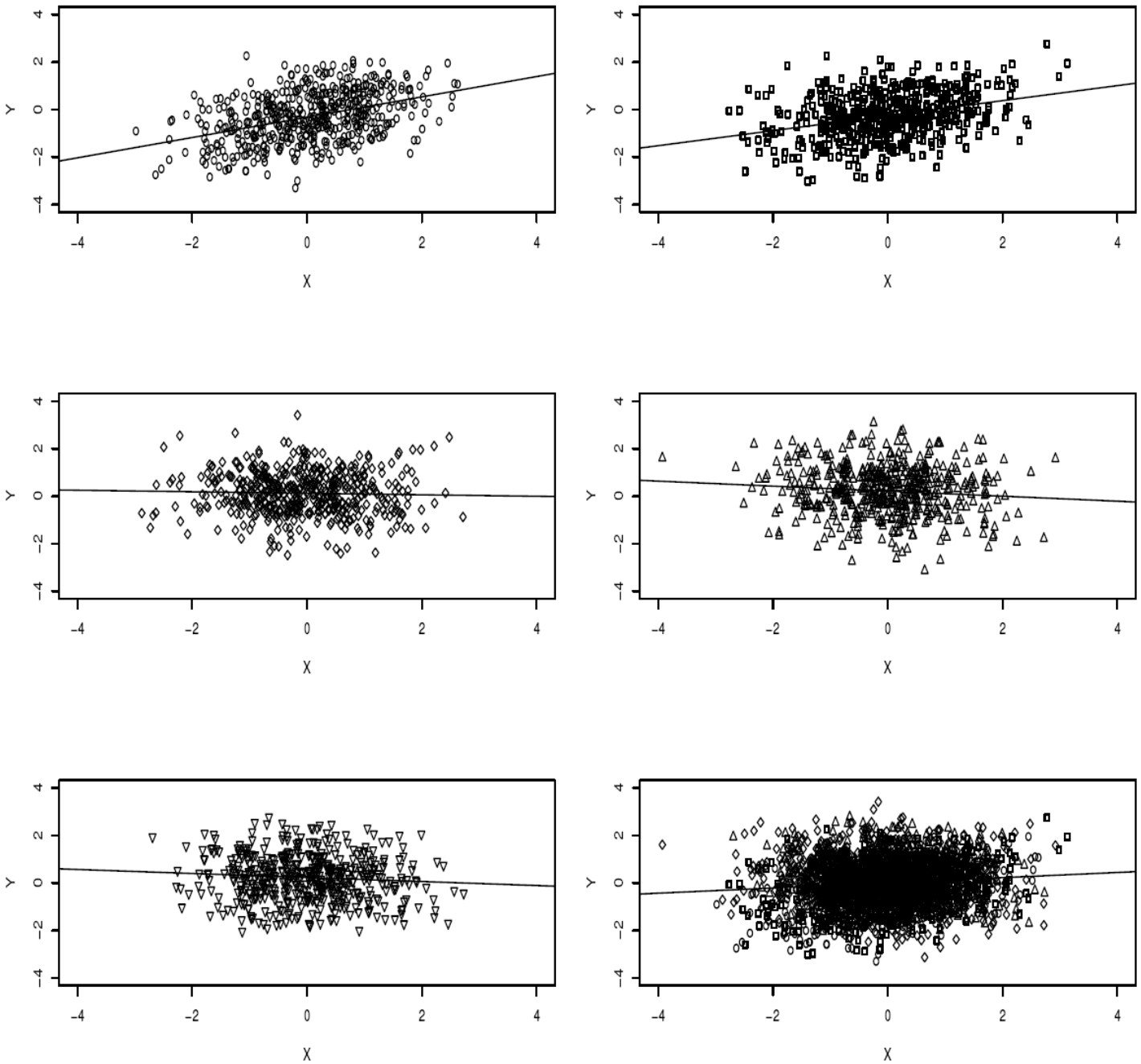


Figure 7.1: Heterogeneity Bias—variable coefficients

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:

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

- Stack the cross-sectional regressions à la Zellner's seemingly unrelated regression model:

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

- 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:

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

- 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 $\mathbf{v}_i \sim N(\mathbf{0}, \boldsymbol{\Gamma})$.

NOTICE that key assumption(s) in line(s) 3(-4) again…

- 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):

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

Again, this from key assumptions 3-4.

- 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:

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

Note the assumption of no cross-unit correlation…

7.3 GLS/FGLS

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

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

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

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

The consistent 1st-stage could be OLS, unit-wise, or SUR, for instance.

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

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

- Using eq. 7.9, we can write:

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

- 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:

$$\hat{\mathbf{V}}_i = s_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$$

$$s_i^2 = \frac{\mathbf{e}_i'\mathbf{e}_i}{T-k}$$

- 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., $\text{var}[\mathbf{b}_i] = \mathbf{V}_i + \hat{\boldsymbol{\beta}}$).

- 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, $\mathbf{b}_i \to \boldsymbol{\beta}_i$, 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

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

where K is some constant.

- 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

New Stata commands allow alternatives to Swamy estimator.

Beck & Katz (2005) MC's of RC in TSCS

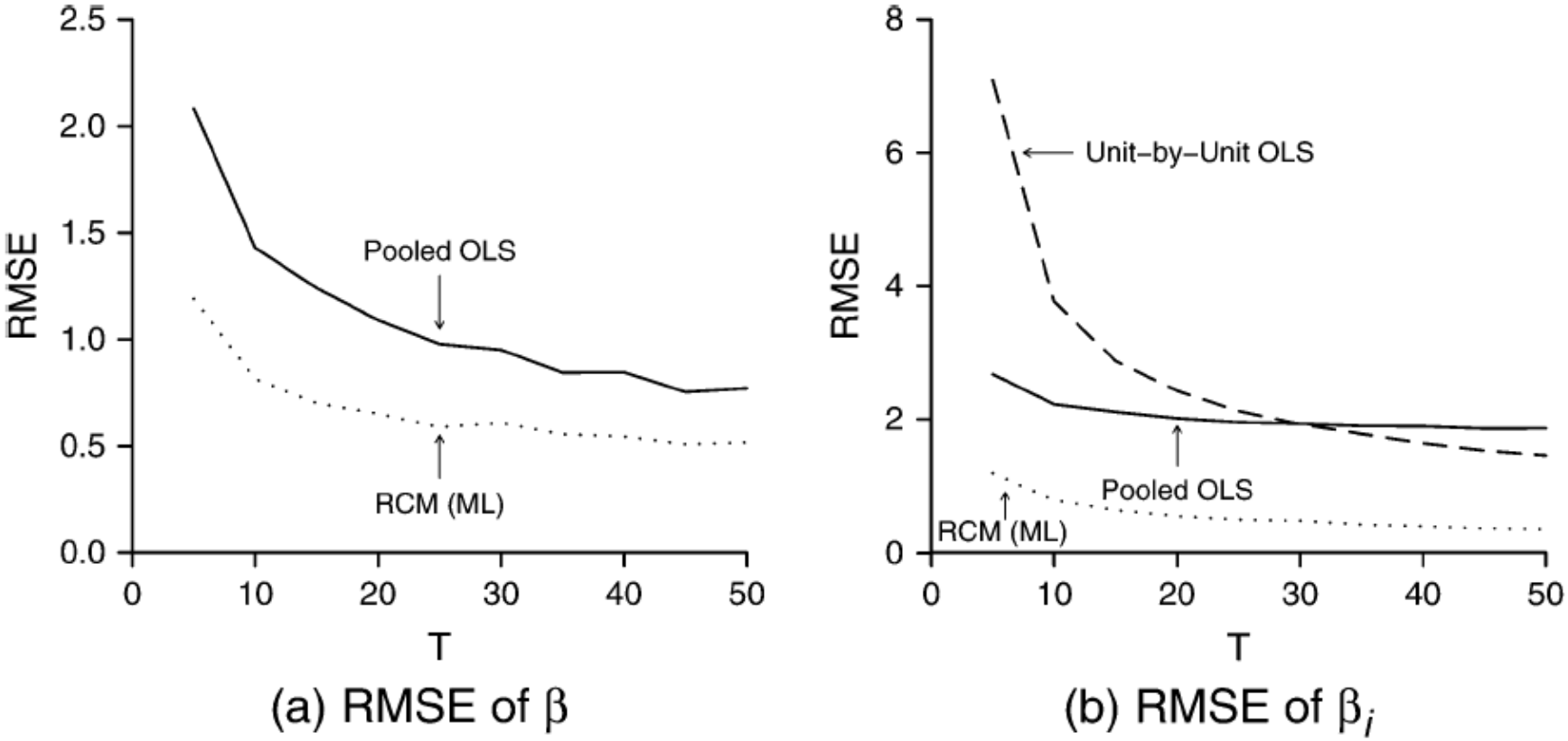


Fig. 2 Comparison of RMSE for RCM and OLS estimators of β and β_i as T varies from 5 to 50. For all runs of the experiment, $N = 20$, $\sigma_\varepsilon^2 = 1$, and $\sigma_x^2 = 0.01$. The β_i were drawn as $\Gamma(5, 1.8^2)$.

Beck & Katz (2005) MC's of RC in TSCS

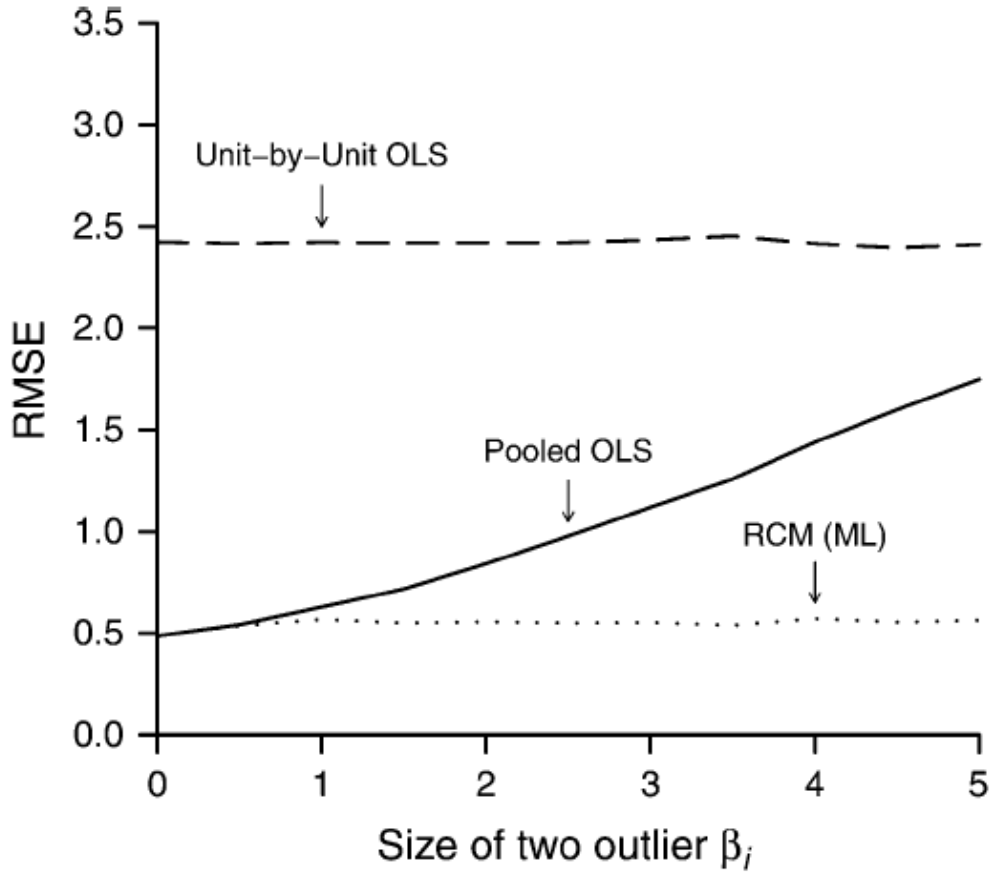


Fig. 3 Comparison of RMSE for RCM and OLS estimators of β_i where the last 18 $\beta_i = 5$ and the first two vary from 5 to 10. For all runs of the experiment, $N = 20$, $T = 20$, $\sigma_\varepsilon^2 = 1$, and $\sigma_x^2 = 0.01$.

Beck & Katz (2005) MC's of RC in TSCS

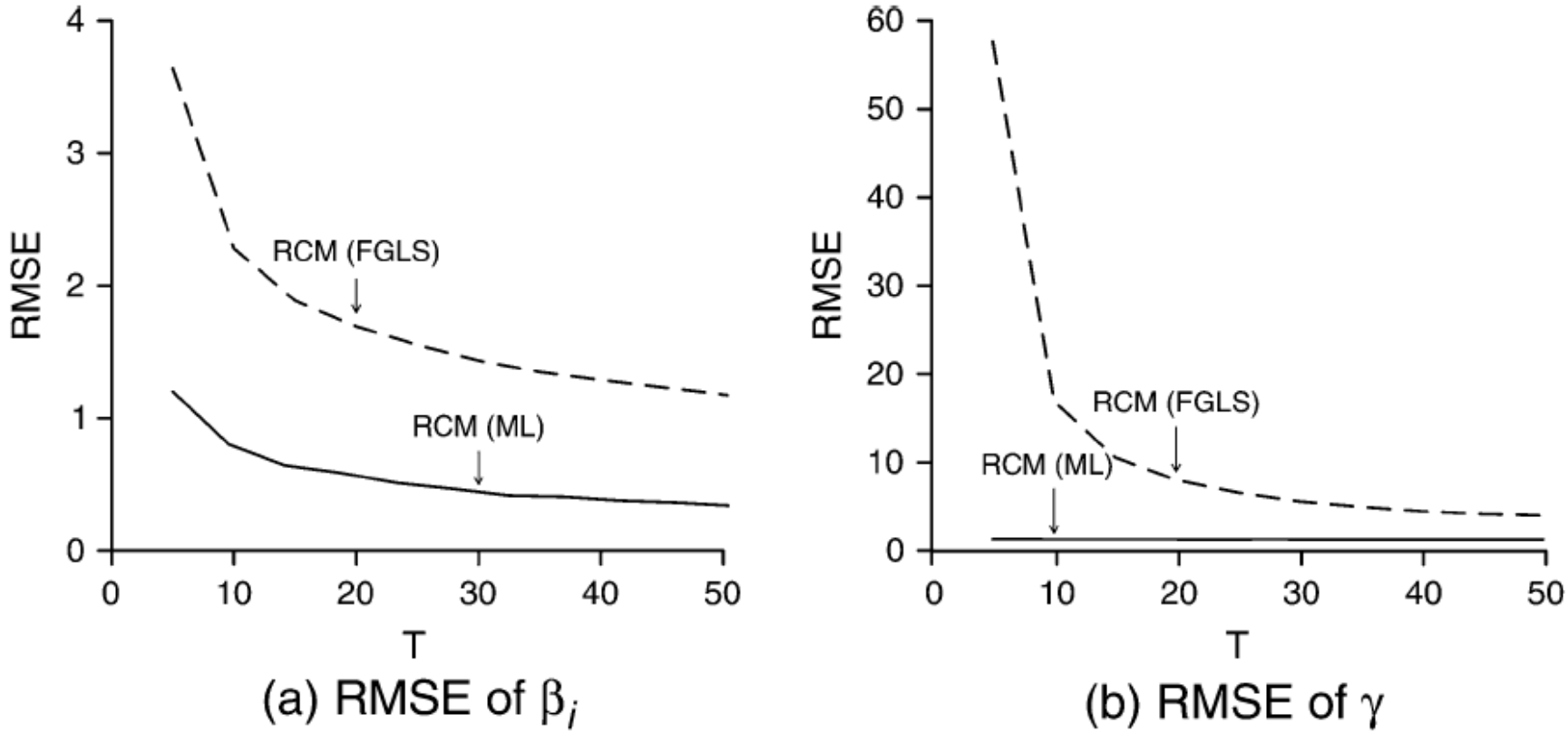


Fig. 4 Comparison of RMSE for RCM and FGLS estimators of β_i and γ as T varies from 5 to 50. For all runs of the experiment, $N = 20$, $\beta = 5$, $\gamma = 1.8$, $\sigma_\varepsilon^2 = 1$, and $\sigma_x^2 = 0.01$.

7.4 Modeling coefficients as functions of exogenous variables

- Our 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 if 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 fully exploit the covariance structure by assuming strict exogeneity of the disturbance terms ε_{ij}, u_{0j} and u_{1j}:

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

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 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.

HLM's (Courtesy of Marco Steenbergen)

Regression Specification

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$$

with $\epsilon_{ij} \sim N(0, \sigma^2)$.

Modified Specification

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij}$$

Intercept and slope are now allowed to vary across level-2 units.

Following Swamy (1970)

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \delta_{0j} \\ \beta_{1j} &= \gamma_{10} + \delta_{1j} \end{aligned}$$

$$\begin{pmatrix} \delta_{0j} \\ \delta_{1j} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix} \right)$$

$$\begin{aligned} y_{ij} &= (\gamma_{00} + \delta_{0j}) + (\gamma_{10} + \delta_{1j})\, x_{ij} + \epsilon_{ij} \\ &= \gamma_{00} + \gamma_{10} x_{ij} + \delta_{0j} + \delta_{1j} x_{ij} + \epsilon_{ij} \\ &= \gamma_{00} + \gamma_{10} x_{ij} + u_{ij} \end{aligned}$$

1. $E[u_{ij}] = 0.$
2. $V[u_{ij}] = \sigma^2 + \tau_{00} + 2\tau_{01} x_{ij} + \tau_{11} x_{ij}^2.$
3. $E[u_{ij}, u_{mj}] = \tau_{00} + \tau_{01}(x_{ij} + x_{mj}) + \tau_{11} x_{ij} x_{mj}.$

Model Elements:

Fixed Effects

γ_{00} and γ_{10}.

Random Effects

δ_{0j}, $\delta_{1j}x_{ij}$, and ϵ_{ij}.

(Co)Variance Components

τ_{00}, τ_{11}, τ_{01}, and σ^2.

Model Assumptions:

1. The level-2 units are exchangeable
2. $E[\delta_{0j}, \epsilon_{ij}] = E[\delta_{1j}, \epsilon_{ij}] = 0$
3. $E[x_{ij}, \delta_{0j}] = E[x_{ij}, \delta_{1j}] = E[x_{ij}, \epsilon_{ij}] = 0$

(i.e., no covariance anywhere; just common error-components within units)

Level-2 Model

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01} z_j + \delta_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} z_j + \delta_{1j}\end{aligned}$$

Combined Model

$$\begin{aligned}y_{ij} &= (\gamma_{00} + \gamma_{01} z_j + \delta_{0j}) + (\gamma_{10} + \gamma_{11} z_j + \delta_{1j}) x_{ij} + \epsilon_{ij} \\ &= \gamma_{00} + \gamma_{01} z_j + \gamma_{10} x_{ij} + \gamma_{11} x_{ij} z_j + \delta_{0j} + \delta_{1j} x_{ij} + \epsilon_{ij}\end{aligned}$$

Causal Heterogeneity

$x_{ij} z_j$ is the cross-level interaction, which captures causal heterogeneity

(There it is again…)

- δ_{0j} and δ_{1j} are now residuals
- τ_{00} and τ_{11} reflect variance in intercepts and slopes unaccounted for by z_j

General, Multivariate HLM

Level-1 Model

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \cdots + \beta_{Pj}x_{Pij} + \epsilon_{ij}$$

Level-2 Model

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}z_{1j} + \gamma_{02}z_{2j} + \cdots + \gamma_{0Q}z_{Qj} + \delta_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}z_{1j} + \gamma_{12}z_{2j} + \cdots + \gamma_{1Q}z_{Qj} + \delta_{1j} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}z_{1j} + \gamma_{22}z_{2j} + \cdots + \gamma_{2Q}z_{Qj} + \delta_{2j}
\end{aligned}$$

etcetera

(Co)Variance Components

1. $E[\epsilon_{ij}^2] = \sigma_j^2$
2. $E[\delta_{pj}^2] = \tau_{pp}$
3. $E[\delta_{pj}, \delta_{qj}] = \tau_{pq}$

- By restricting γ_{pq} we can influence which level-2 covariates influence the level-1 effects
- When $\beta_{pj} = \gamma_{p0}$ this is tantamount to saying that there is no variation in an effect across level-2 units

Let

- $\mathbf{y}$ be a vector containing the response variable
- $\mathbf{X}$ be the matrix of level-1 covariates
- $\mathbf{Z}$ be the matrix of level-2 covariates
- $\boldsymbol{\delta}$ be a vector containing the level-2 errors
- $\boldsymbol{\epsilon}$ be a vector containing the level-1 errors
- $\mathbf{W}$ be the matrix of level-2 variance components
- $\boldsymbol{\Omega}$ be the matrix of level-1 variance components

Then $\cdots$

Structural Model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\mathbf{Z}\gamma + \mathbf{X}\delta + \epsilon \\ &= \mathbf{X}\mathbf{Z}\gamma + \mathbf{u} \end{aligned}$$

Variance Model

$$\begin{aligned} \mathbf{V} &= E[\mathbf{u}\mathbf{u}'] \\ &= \mathbf{X}\left(\mathbf{W} \otimes \mathbf{I}_J\right)\mathbf{X}' + \Omega \end{aligned}$$

(I.e., a special case of the GLRM…)

Interpretation looks familiar also.

Consider classic interaction model:

Model

$$\begin{aligned} y_{ij} &= \gamma_{00} + \gamma_{01} z_j + \gamma_{10} x_{ij} + \gamma_{11} z_j x_{ij} + \\ &\quad \delta_{0j} + x_{ij}\delta_j + \epsilon_{ij} \end{aligned}$$

Simple Effect of x_{ij}

$$\frac{\partial y_{ij}}{\partial x_{ij}} = \gamma_{10} + \gamma_{11} z_j + \delta_{0j}$$

Marginal Effect of the Interaction

$$\frac{\partial^2 y_{ij}}{\partial x_{ij} \partial z_j} = \gamma_{11}$$

Simple Effect of z_j

$$\frac{\partial y_{ij}}{\partial z_j} = \gamma_{01} + \gamma_{11} x_{ij}$$

Marginal Effect of the Interaction

$$\frac{\partial^2 y_{ij}}{\partial z_j \partial x_{ij}} = \gamma_{11}$$

Estimation:

ML (Restricted & Penalized ML also exist)

Full Information ML

$$\begin{aligned}\ell &= -.5\left[n\ln 2\pi + \ln|\mathbf{V}| + (\mathbf{y} - \mathbf{XZ}\gamma)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{XZ}\gamma)\right] \\ &= -.5\left[n\ln 2\pi + \ln|\mathbf{V}| + \mathbf{u}'\mathbf{V}^{-1}\mathbf{u}\right]\end{aligned}$$

Algorithms

- EM algorithm (HLM; see Raudenbush & Bryk 1986)
- Fisher scoring (VARCL; see Longford 1987)
- IGLS (MLwiN; see Goldstein 1986)
- In the normal case, these are equivalent

Shrinkage Estimation:

EB/Shrinkage Estimation

$$\tilde{\boldsymbol{\beta}}_j = \mathbf{Q}_j \hat{\boldsymbol{\beta}}_j + (\mathbf{I} - \mathbf{Q}_j) \breve{\boldsymbol{\beta}}_j$$

- $\hat{\boldsymbol{\beta}}_j$ is the within-group OLS estimator
- $\breve{\boldsymbol{\beta}}_j = \mathbf{Z}_j \hat{\gamma}$ is the pooled estimator
- $\mathbf{Q}_j = \mathbf{W}\mathbf{D}_j^{-1}$ with $\mathbf{D}_j = \mathbf{W} + \sigma_j^2 (\mathbf{X}_j' \mathbf{X}_j)^{-1}$

Bayesian Estimation:

Theorem

$$g(\gamma, \mathbf{V} | Data) \propto L(Data | \gamma, \mathbf{V}) f(\gamma, \mathbf{V})$$

Legend

- $f(\gamma, \mathbf{V})$ is the prior
- $L(Data | \gamma, \mathbf{V})$ is the likelihood
- $g(\gamma, \mathbf{V} | Data)$ is the posterior

For all this, as we've seen, not much new in interpretation & testing…

Temporally Dynamic Panel-Data Models
(Wawro again)

- 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.

The Model:

- 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 i denotes the cross-sectional units ($i = 1, \ldots, N$), t denotes the time period ($t = 1, \ldots, T$), $x_{i,t}$ is an exogenous explanatory variable, γ and β are parameters to be estimated, α_i is an individual-specific effect, and $u_{i,t}$ is a random disturbance term.

- 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.

OLS Biased if Inadequate Model Unit-Specific

- 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.

I called this inverse Hurwicz/Nickell before. It's just another example of OVB.

LSDV/FE LDV is Small-(*T*-)Sample Biased

- 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.

In fact, bias is "of order 1/*T*", which means proportionate bias is 1/*T*. (So 5% for T=20, e.g., but 25% for *T*=4.)
(So big-to-huge issue for Panel, moderate-to-small issue for TSCS.)

Panel-analyst's response to conundrum:
Model unit-effects, then redress Hurwicz Bias.

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

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

  where Δ is the difference operator such that the notation $\Delta y_{i,t} = y_{i,t} - y_{i,t-1}$.

- 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 $\Delta u_{i,t}$.

Instruments from TSCS Structure Dataset…

- 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:

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

Anderson-Hsiao Insight & Problems

- 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.

"Improvements": GMM Estimators - i.e., (Asymptotic-)Efficiency Enhancers (if true)

- 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 vari-
- More generally, GMM : 2SLS : : 2SLS : IV – namely, uses more info & *id* restricts
- Familiar "exclusion restrictions" regard 1st moment; extra GMM ones regard 2nd (&, in principle, higher) moments.
- In implementation, can show GMM is matrix-wtd 2SLS. So could also say GMM : 2SLS : :GLS : OLS

"Review" GMM Estimation (1)

- 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.

MoM: use sample analog as estimator population parameter. So,

e.g.: $\bar{x}$ is a MoM estimator for μ_x.

G-MoM: use higher moment conditions too.

So, G-MoM : MoM : : GLS : OLS

"Review" GMM Estimation (2)

- 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.

OLS is also a MoM, in other words...

- 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 $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ 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.

Instrumental Variables in GMM Framework

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

Just identified = IV; Overidentified = 2SLS

Capitalizing on the Overidentifying Information

- To get around this problem we choose $\boldsymbol{\beta}$ 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

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

Optimizing the Weights in GMM

- 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 $\boldsymbol{\Omega}$ 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

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

- 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.

GMM for (Temporally) Dynamic Panel-Data

Arrellano-Bond enhance Anderson-Hsiao

- 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

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

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

- For notational efficiency, we can stack the time periods to write dowı 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

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

- Note that this means that the number of instruments increases as move through the panel. For example, if we have the simple mo 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

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

Issue in 2SLS/GMM to keep rank($\mathbf{Z}$) not too large relative to rank(endog)…b/c overfitting 1st-stage => fitting some of endog.

- 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

- 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.

- 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*).

<u>TO RESTATE</u>: Properties are Asymptotic, <u>*in N*</u>

Twin Horns of Dilemma for Estimator:

- Instruments too weak, Instruments too many.

6.6.2 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.

- Cottage Industry of Further Enhancements:

- Changes as well as levels (& v.v.) equally valid instruments.
  - "Orthogonal Deviations" (actually forward obs: leads)
    - "Robustness" to non-spherical V-Cov

6.6.3 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.

AB's Omnibus Over-ID Test:

- 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 eliminate 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.

(Help xtdpdsys)

[The above should be a link to PDF of PowerPoint Presentation on Models of Spatial Interdependence]

QualDep Models in TSCS

XV. Introduction

A. TSCS considerations arise for limited/qualitative-dependent-variable (*QualDep*) models also…

B. Non-separability of stochastic from systematic components in these models renders the proper address of these considerations considerably more complicated.

C. Methods have been developed for binary, polychotomous, censored, truncated, ordered, count, duration, etc. models.

D. *Stimson's Law*: You can only solve one hard problem at a time, and solving it requires ignoring lots of other problems.

XVI. Binary Dependent-Variable Models

$$y^*_{i,t} = \mathbf{x}_{i,t}\beta + \epsilon_{i,t}$$

$$y_{i,t} = 1 \text{ if } y_{i,t} > 0$$

A. Typical Procedure for ML Analysis:

1. Distribution of *DepVar*; Bernouli: $\Pr(y_{it} = \{1,0\}) = p_{it}^{y_{it}} \times (1 - p_{it})^{1-y_{it}}$

2. Model parameters of interest; includes appropriate covariates, **X**, and appropriate functional form ("link function"). For binary (draw):

   a) Sigmoidal (S-Shaped) link, logit: $p = e^{\mathbf{x\beta}}\left(1 - e^{\mathbf{x\beta}}\right)^{-1}$

   b) Sigmoidal (S-Shaped) link, probit: $p = \Phi\left(\mathbf{x\beta}\right)$

   c) Many, many other sigmoidal functions; some useful ones relax symmetry, steepest at 0.5, etc.

3. Ensure—theoretically/substantively, by good/clever specification of **xβ**—or assume (or pray for)—conditional independence of obs, so

4. Joint likelihood of obs given model & data is product marginals; maximize log-likelihood for parameter estimates; $-\mathbf{H}^{-1}$ are vce.

B. Binary Models with Unit-Specific Effects

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.
- Unfortunately, the inconsistency in α_i is transmitted to $\boldsymbol{\beta}$.
- 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.

1. Fixed-Effect Logit (a.k.a., Conditional Logit):

a) Chamberlin's *Clever strategy/insight*: Conditioning on the number of ones, i.e., on $\sum_{t=1}^{T} y_{it}$, works like fixed-effect transformation for binary, i.e., like FE/LSDV, because the FE are conditional means, and, in this context, those are probabilities, which are sums of y_{it}=1 (divided by T). Formally: $\sum_{t=1}^{T} y_{it}$ is a sufficient statistic for α_i.

There is no simple method for fixed effects binary panel data. The problem is the Neyman-Scott incidental parameter problem discussed on Tuesday. Because the probit/logit model is non-linear, there is no nice way to sweep out the unit effects, and inconsistencies ir the unit effects then cause inconsistent estimation of β. Some analyses show that for the inconsistency is $O(\frac{1}{T}$ and is about 50% for $T = 2$.

(1) That is, simple FE is biased in "sigmoidal Hurwicz/Nickell" fashion.

(2) Implications wider-ranging b/c sigmoidal function non-separability.

(3) However, bias also like Hurwicz/Nickell in that of order 1/T, meaning in reasonably long TSCS, may not be an issue worth fretting.

b) Formally, the likelihood for Chamberlin's Conditional Logit is:

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

c) How it works/what it does, is easiest explained & seen in T=2 case:

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

- 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

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

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

d) Where the second denominator equals the first because the other two cases (both observations =1) has been discarded (conditioned away).

- 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

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

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

e) …is essentially a multinomial-logit form.

f) The other expression is: $e^{\mathbf{x}_{i1}\boldsymbol{\beta}} / (e^{\mathbf{x}_{i1}\boldsymbol{\beta}} + e^{\mathbf{x}_{i2}\boldsymbol{\beta}})$

g) The likelihood to maximize takes these (rel'ly familiar) MNL forms.

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

h) i.e., number of orderings of 0's & 1's increases combinatorically in T.

i) Note: this at least analogous if not identical to the strategy & issues in network ERGM's (Exponential Random Graph Models).

j) Wawro's Commentary:

- 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:

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

  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.

k) Beck's Commentary:

Note we are conditioning on the number of successes (1's) for unit i, so there is a lot of conditioning going on here.

The basic idea is that the α determines the overall proportion of successes in any unit, and the β and $\mathbf{x}$ determine in which years of unit i the successes are most likely.

We have already seen that if a unit has two negative outcomes, we just α_i as large negative as possible and we get no information on β. Same for two positive outcomes (make α_i as large as possible). What is going on is that the conditional approach only gets information about β for units with some failures and some successes, with that information being the conditional probability of a success given the number of successes.

(Note that if this is true, then conditional logit can give us little information on covariates that change slowly. Take democracy, for example. Any unit where democracy is stable gives us no information on the effect of democracy, since this effect must be the same in the failures and successes in that unit, and any cross unit differences are accounted for by the α, not the covariates.)

NOW IF WE ASSUME THE P's ARE GENERATED STANDARD LOGIT, THIS SIMPLIFIES NICELY AND THE EFFECTS DISAPPEAR

Using the logit form, it is quite easy to write down these joint probabilities (done on Green p. 840), the α_i drop out of this equation and the conditional probability of the sequence (0,1) is just a logit. Nothing deep here - if you know you had one success out of two trials, the only information in the data about β is given by whether the first or second trial was positive, with that probability given to you by a logit BASED ONLY ON THE TWO OBSERVATIONS FOR UNIT i. (Thus conditional logit works for the same reason that with more than two outcomes you can do logit if you assume IIA.)

Note that conditional fixed effects logit is not exactly fixed effects logit, its qualities are asymptotic, and if you have a lot of units with all zeros or all one, you are in trouble. If you like fixed effects for continuous dv's, then this procedure inherits the good from that; if you don't like continuous fe's, it inherits the bad.

Note also that Ethan Katz has shown that as $T \rightarrow \infty$ that standard logit with fixed effects and conditional logit converge (this is well known), but in reality they are very close when $T > 15$. Thus only need Chamberlain for smallish T, for largish T just put in dummy variables for unit and run logit.

And of course, for fe logit, you have to believe that all the information in the data about the β, the parameters of interest, is contained in *when* observations in a unit are zero or one, *not* how many are zero or one. This seems to be throwing a lot of information away (just as in the discussion of Green, Kim and Yoon for dyadic BTSCS data).

2. Random-Effect Probit:

a) In probit, as Beck alluded, conditioning on fixed-effect $\sum_{t=1}^{T} y_{it}$ does not yield a familiar simplified expression (like C/FE-Logit's MNL form). Interestingly, though, RE simplifies better/more/more-easily in probit.

$$y_{it}^{*} = \mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it} \text{ , with } \alpha_i \sim N\left(0, \sigma_{\alpha}^{2}\right)$$

$$y_{it} = \left\{1 \text{ if } y_{it}^{*} > 0 \text{ ; } 0 \text{ if } y_{it}^{*} < 0\right\}$$

b) Excluding the RE, we would just apply standard-ML probit, multiplying the *NT* marginal likelihoods to obtain the joint likelihood.

c) With random-effects—i.e., with the α_i being random variables—the joint likelihood is the product *N* inseparable *T*-dimensional likelihoods.

d) Note: this is same issue that arises in spatial (& multilevel) *QualDep* models also. The following strategy not work in space, though.

e) That's very hard (computationally intensive) to integrate. Applicable trick here (popularized by Butler and Moffitt, but known before) is to condition on α_i, which makes the *T* densities independent:

Writing $\nu_{i,t} = \alpha_i + \epsilon_{i,t}$, we then get

$$f(\nu_{i,1}, \nu_{i,2}, \ldots, \nu_{i,T}) = \int_{-\infty}^{\infty} f(\nu_{i,1}, \nu_{i,2}, \ldots, \nu_{i,T} | \alpha_i) f(\alpha_i) d\alpha_i$$

$$= \int_{-\infty}^{\infty} \prod_{t=1}^{T} f(\nu_{i,t} | \alpha_i) f(\alpha_i) d\alpha_i$$

f) This now just a product of 1-dimensional marginals, so feasible, but to get there, integrate over the α_i, so still rather intense. `xtprobit`, much slower than `probit`, exponentially so in T. T>10? impractical.

g) As usual, the REs must be <u>*assumed*</u> independent of **x**'s. Wawro *AJPS* (2001) discusses a correlated-RE logit. It makes some odd/strong assumptions of its own, of course, and is very hard to estimate.

Wawro offers a fuller discussion:

- 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

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

h) First, express the joint as conditionals times marginal (Bayes Law):

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

i) Then integrate over the 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:

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

j) Now, write the univariate normals in the product and express as $g(\alpha_i)$:

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

k) Then we can change some notation and rearrange and get this into the form the computer actually uses to search for parameter estimates:

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

Wawro's Commentary:

- Things to note:
  - ➢ The assumption that the α_i and $\mathbf{x}_{it}$ are uncorrelated is very restrictive. 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.

8.8 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.

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

Note: BKT time-dependence in BTSCS by time dummy/splines is a kludge…

Lagged DV vs Lagged Latent Models

Or one could do ML (easiest wit MCMC, see various Jackman pieces) to estimate one of three models in the latent y^*:

$$y^*_{i,t} = \mathbf{x}_{i,t}\beta + \epsilon_{i,t} + \rho\epsilon_{i,t-1} \tag{19}$$

$$y^*_{i,t} = \mathbf{x}_{i,t}\beta + \epsilon_{i,t} + \rho y_{i,t-1} \tag{20}$$

$$y^*_{i,t} = \mathbf{x}_{i,t}\beta + \epsilon_{i,t} + \rho y^*_{i,t-1} \tag{21}$$

Equation 19 is just like an AR1 error model (though it is actually MA1 error, hard to tell apart and easier to notate!); Equation 20 is a model of "true" state dependence and Equation 21 is call "spurious" state dependence.

Note the difference - in true state dependence what matters is the realized dv (going to war makes you more likely to go to war next year, being employed this month makes you more likely to be employed next month), spurious state dependence has the underlying propensity to go to war persist, doesn't matter if you actually get the war.

Only true state dependence model is easy to estimate, just throw lagged y into specification.

Others are doable, though. Like spatial-probit, except somewhat easier estimate because **W** for time, our **V**, is lower-triangular.