Applied Regression Models for Longitudinal Data

Kosuke Imai

Princeton University

Fall 2016
POL 573 Quantitative Analysis III

Readings

- Hayashi, *Econometrics*, Chapter 5
- "Dirty Pool" papers referenced in the slides
- Wooldrich, *Econometric Analysis of Cross Section and Panel Data*, Chapter 10 and relevant sections of Part IV
- Gelman and Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Part 2A, Cambridge University Press.

What Are Longitudinal Data?

- Repeated observations for each unit
- Also called panel data, cross-section time-series data
- Assume *balanced* data:

$$\underbrace{Y_i}_{T \times 1} = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{pmatrix} \quad \text{and} \quad \underbrace{X_i}_{T \times K} = \begin{pmatrix} x_{i11} & x_{i12} & \cdots & x_{i1K} \\ x_{i21} & x_{i22} & \cdots & x_{i2K} \\ \vdots & \cdots & \cdots & \vdots \\ x_{iT1} & x_{iT2} & \cdots & x_{iTK} \end{pmatrix}$$

- "Stacked" form:

$$\underbrace{\mathbf{Y}}_{NT \times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} \quad \text{and} \quad \underbrace{\mathbf{X}}_{NT \times K} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix}$$

Varying Intercept Models

- Basic setup:

$$y_{it} = \alpha_i + \beta^\top x_{it} + \epsilon_{it}$$

- Motivation: unobserved (time-invariant) heterogeneity

$$y_{it} = \alpha + \beta^\top x_{it} + \delta^\top u_i + \epsilon_{it}$$

  where $\alpha_i = \alpha + \delta^\top u_i$
- (Strict) Exogeneity given X_i and u_i:

$$\mathbb{E}(\epsilon_{it} \mid X_i, u_i) = 0$$

- Constant slopes
- Static model: no lagged y in the right hand-side

Fixed-Effects Model

- "Fixed" effects mean that α_i are model parameters to be estimated
- $(N + K)$ parameters to be estimated: inefficient if T is small
- Homoskedasticity and independence across time (& units)

$$\mathbb{V}(\epsilon_i \mid \mathbf{X}) \; = \; \sigma^2 I_T$$

- Stacked vector representation:

$$\mathbf{Y} \; = \; \mathbf{Z}\gamma + \epsilon \text{ where } \mathbf{D} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \; \mathbf{Z} = [\mathbf{D}\,\mathbf{X}], \; \gamma = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \\ \beta \end{pmatrix}$$

Estimation of Fixed Effects Model

- The least-squares estimator (also MLE): $\hat{\gamma}_{FE} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{Y}$
- Sampling distribution: $\hat{\gamma}_{FE} \mid \mathbf{Z} \sim \mathcal{N}(\gamma, \sigma^2 (\mathbf{Z}^\top \mathbf{Z})^{-1})$
- The dimension of $(\mathbf{Z}^\top \mathbf{Z})$ is large when N is large
- Computation based on within-group variation:

$$\widetilde{\mathbf{Y}} = \begin{pmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{iT} - \bar{y}_i \\ \vdots \\ y_{NT} - \bar{y}_N \end{pmatrix} \quad \text{and} \quad \widetilde{\mathbf{X}} = \begin{pmatrix} (x_{11} - \bar{x}_1)^\top \\ \vdots \\ (x_{iT} - \bar{x}_i)^\top \\ \vdots \\ (x_{NT} - \bar{x}_N)^\top \end{pmatrix}$$

- Then, $\hat{\beta}_W = (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{Y}}$ and $\hat{\beta}_W \mid \mathbf{X} \sim \mathcal{N}(\beta, \sigma^2 (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1})$

Fixed Effects Estimator as Within-group Estimator

- Within-group variation $=$ Residuals from regression of $\mathbf{Y}$ on $\mathbf{D}$
- Recall the geometry of least squares (see Multiple Regression slides 39 and 43)
- Projection of $\mathbf{Y}$ onto $\mathcal{S}^{\perp}(\mathbf{D})$
- Thus, $\widetilde{\mathbf{Y}} = \mathbf{M_D Y} =$ where $\mathbf{M_D} = \mathbf{I}_{NT} - \mathbf{D}(\mathbf{D}^\top \mathbf{D})^{-1}\mathbf{D}^\top$
- Also, $\widetilde{\mathbf{X}} = \mathbf{M_D X}$
- This is a partitioned regression!
- Then, $\hat{\beta}_{FE} = \hat{\beta}_W$ (see Question 1 of POL572 Problem Set 4 ☺)
- Also, $\hat{\epsilon} = \mathbf{Y} - \mathbf{Z}\hat{\gamma}_{FE} = \widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\hat{\beta}_W$
- The lower-right block of $(\mathbf{Z}^\top \mathbf{Z})^{-1}$ equals $(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1}$
- Thus, $\hat{\beta}_W \mid \mathbf{X} \sim \mathcal{N}(\beta, \sigma^2 (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1})$

Kronecker Product

- Definition:

$$\mathbf{A}_{n\times m} \otimes \mathbf{B}_{k\times l} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1m}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & a_{n2}\mathbf{B} & \dots & a_{nm}\mathbf{B} \end{bmatrix}_{nk\times ml}$$

- Some rules (assuming they are conformable):
  - $(\mathbf{A} \otimes \mathbf{B})^\top = \mathbf{A}^\top \otimes \mathbf{B}^\top$
  - $\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}$
  - $(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$
  - $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$
  - $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$
  - $|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^m |\mathbf{B}|^n$ where $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ and $m \times m$ matrices, respectively.
- Fixed effects calculation made easy: $\mathbf{D} = \mathbf{I}_N \otimes \mathbf{1}_T$ which implies
  $\mathbf{D}^\top \mathbf{D} = \mathbf{I}_N \otimes (\mathbf{1}_T^\top \mathbf{1}_T) = T\mathbf{I}_N$
  $\mathbf{P}_\mathbf{D} = \mathbf{D}(\mathbf{D}^\top \mathbf{D})^{-1}\mathbf{D}^\top = \frac{1}{T}\mathbf{I}_N \otimes (\mathbf{1}_T \mathbf{1}_T^\top)$

Serial Correlation and Heteroskedasticity

- Even after conditioning on x_{it} and u_i, y_{it} may still be serially correlated
- $\text{Corr}(\epsilon_{it}, \epsilon_{it'}) \neq 0$ for $t \neq t'$
- Independence across units is assumed
- We are still assuming we got the conditional mean specification correct and so just need to "fix" standard errors
- Robust standard errors (see Multiple Regression slide 24):

$$\mathbb{V}(\hat{\beta} \mid \mathbf{X}) = \underbrace{(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1}}_{bread} \underbrace{\{\widetilde{\mathbf{X}} \mathbb{E}(\tilde{\epsilon}\tilde{\epsilon}^\top \mid \mathbf{X}) \widetilde{\mathbf{X}}\}}_{meat} \underbrace{(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1}}_{bread}$$

  where the "meat" is estimated by $\sum_{i=1}^N \widetilde{X}_i^\top \hat{\epsilon}_i \hat{\epsilon}_i^\top \widetilde{X}_i$
- Asymptotically consistent with any form of heteroskedasticity and serial correlation
- Can also use Feasible GLS

Panel Corrected Standard Error

- A popular procedure proposed by Beck and Katz (1995)
- Basic idea: take into account contemporaneous (or spatial) correlation when calculating standard errors
- Autocorrelation is assumed to be non-existent

$$\mathbb{E}(\tilde{\epsilon}_{it}\tilde{\epsilon}_{it'} \mid \mathbf{X}) = \mathbb{E}(\tilde{\epsilon}_{it}\tilde{\epsilon}_{i't'} \mid \mathbf{X}) = 0 \quad \text{for } i \neq i' \text{ and } t \neq t'$$

- Inclusion of lagged dependent variable (more on this later)
- Spatial correlation is assumed to be time-invariant

$$\hat{\rho}_{ii'} \;=\; \widehat{\mathbb{E}(\tilde{\epsilon}_{it}\tilde{\epsilon}_{i't} \mid \mathbf{X})} \;=\; \frac{1}{T}\sum_{t'=1}^{T}\hat{\epsilon}_{it'}\hat{\epsilon}_{i't'} \quad \text{for } i \neq i'$$

- These robust standard errors can be applied to any linear models (with or without fixed effects)

A Variety of Exogeneity Assumptions

- Contemporaneous exogeneity: $\mathbb{E}(\epsilon_{it} \mid x_{it}, \alpha_i) = 0$
- Not sufficient for identification of β under fixed effects model

- Strict exogeneity: $\mathbb{E}(\epsilon_{it} \mid X_i, \alpha_i) = 0$
- Sufficient for identification of β under fixed effects model
- error does not correlate with x at another time

- Sequential exogeneity: $\mathbb{E}(\epsilon_{it} \mid \overline{X}_{it}, \alpha_i) = 0$ where $\overline{X}_{it} = \{x_{i1}, x_{i2}, \ldots, x_{it}\}$
- past error can correlate with future x

- Dynamic sequential exogeneity: $\mathbb{E}(\epsilon_{it} \mid \overline{X}_{it}, \overline{Y}_{i,t-1}) = 0$ where $\overline{Y}_{it} = \{y_{i0}, y_{i1}, \ldots, y_{it}\}$
- analogous to sequential ignorability

- Sequential ignorability: $\{y_{it}(1), y_{it}(0)\} \perp\!\!\!\perp x_{it} \mid \overline{X}_{i,t-1}, \overline{Y}_{i,t-1}$
- sequential randomization, marginal structural models

Fixed Effects and Lagged Dependent Variable

- One of the simplest fixed effects model that incorporates dynamics is the following AR(1) model:

$$y_{it} = \alpha_i + \rho y_{i,t-1} + \beta^\top x_{it} + \epsilon_{it} \quad \text{where} \quad |\rho| < 1$$

- Strict exogeneity does not hold: $\mathbb{E}(\epsilon_{it} \mid \overline{X}_{iT}, \overline{Y}_{i,T-1}, \alpha_i) \neq 0$
- Bias (Nickell) as N goes to ∞ with fixed T:

$$\begin{aligned}
\hat{\rho} - \rho &= \underbrace{\left(\frac{1}{NT}\widetilde{\mathbf{Y}}_{-1}^\top \mathbf{M}_{\widetilde{\mathbf{X}}} \widetilde{\mathbf{Y}}_{-1}\right)}_{\mathbf{A}_T}^{-1} \frac{1}{NT}\widetilde{\mathbf{Y}}^\top \tilde{\epsilon} \\
&\xrightarrow{p} -\mathbf{A}_T^{-1} \cdot \frac{\sigma^2}{T(1-\rho)}\left(1 - \frac{1-\rho^T}{T(1-\rho)}\right) \\
\hat{\beta} - \beta &= (\rho - \hat{\rho})(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{Y}}_{-1} + (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top \tilde{\epsilon} \\
&\xrightarrow{p} -\mathbf{A}_T^{-1} \cdot \frac{\sigma^2}{T(1-\rho)}\left(1 - \frac{1-\rho^T}{T(1-\rho)}\right) \cdot \delta
\end{aligned}$$

where $(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{Y}}_{-1} \xrightarrow{p} \delta$

First Differencing for Identification

- The model: $y_{it} = \alpha_i + \rho y_{i,t-1} + \beta^\top x_{it} + \epsilon_{it}$
- Assumption: dynamic sequential exogeneity $\mathbb{E}(\epsilon_{it} \mid \overline{X}_{it}, \overline{Y}_{i,t-1}) = 0$
- Implies $\mathbb{E}(\epsilon_{it}) = \mathbb{E}(\epsilon_{it}\epsilon_{it'}) = 0$ for $t \neq t'$
- First differencing:

$$\Delta y_{it} \;=\; \rho \Delta y_{i,t-1} + \beta^\top \Delta x_{it} + \Delta \epsilon_{it}$$

  where $\Delta y_{it} = y_{it} - y_{i,t-1}$ etc.
- Instrumental variables:
  - Anderson and Hsiao: $y_{i,t-2}$ or $\Delta y_{i,t-2}$
  - Arellano and Bond: use all instruments with GMM
    - $t = 3$: $\mathbb{E}(\Delta\epsilon_{i3} y_{i1}) = 0$
    - $t = 4$: $\mathbb{E}(\Delta\epsilon_{i4} y_{i2}) = \mathbb{E}(\Delta\epsilon_{i4} y_{i1}) = 0$
    - $t = 5$: $\mathbb{E}(\Delta\epsilon_{i5} y_{i3}) = \mathbb{E}(\Delta\epsilon_{i5} y_{i2}) = \mathbb{E}(\Delta\epsilon_{i5} y_{i1}) = 0$
    - and so on; a total of $(T-2)(T-1)/2$ instruments
- Instruments derived from the model rather than from the qualitative information

Random Effects Model

- Dimension reduction is desirable when N is large relative to T
- "Random" intercepts: a prior distribution on α_i

$$\alpha_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\alpha, \omega^2)$$

- Additional assumptions:
  1. A family of distributions for α_i
  2. Independence between α_i and $\mathbf{X}$
- Reduced form when $\epsilon_{it} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$:

$$Y_i \mid \mathbf{X} \stackrel{\text{indep.}}{\sim} \mathcal{N}(\alpha \mathbf{1}_T + X_i\beta, \ \sigma^2\Sigma_\tau) \quad \text{where} \quad \Sigma_\tau = \mathbf{I}_T + \tau \mathbf{1}_T \mathbf{1}_T^\top$$

and $\tau = \omega^2/\sigma^2$ or using the stacked form

$$\mathbf{Y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{Z}\gamma, \ \sigma^2\Omega_\tau) \quad \text{where} \quad \Omega_\tau = \mathbf{I}_N \otimes \Sigma_\tau$$

and $\mathbf{Z} = [\mathbf{1} \ \mathbf{X}]$ and $\gamma = (\alpha, \beta)$

Maximum Likelihood Estimation of RE Model

- Likelihood function:

$$(2\pi)^{-NT/2}|\sigma^2\Omega_\tau|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{Z}\gamma)^\top\Omega_\tau^{-1}(\mathbf{Y}-\mathbf{Z}\gamma)\right\}$$

- The same trick as in linear regression:

$$\begin{aligned}
& (\mathbf{Y}-\mathbf{Z}\gamma)^\top\Omega_\tau^{-1}(\mathbf{Y}-\mathbf{Z}\gamma) \\
= \; & (\mathbf{Y}-\mathbf{Z}\gamma-\mathbf{Z}\hat{\gamma}+\mathbf{Z}\hat{\gamma})^\top\Omega_\tau^{-1}(\mathbf{Y}-\mathbf{Z}\gamma-\mathbf{Z}\hat{\gamma}+\mathbf{Z}\hat{\gamma}) \\
= \; & (\mathbf{Y}-\mathbf{Z}\hat{\gamma})^\top\Omega_\tau^{-1}(\mathbf{Y}-\mathbf{Z}\hat{\gamma})+(\gamma-\hat{\gamma})^\top\mathbf{Z}^\top\Omega_\tau^{-1}\mathbf{Z}(\gamma-\hat{\gamma})
\end{aligned}$$

  where $\hat{\gamma}=(\mathbf{Z}^\top\Omega_\tau^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top\Omega_\tau^{-1}\mathbf{Y}$

- Log-likelihood function:

$$-\frac{TN}{2}\log(2\pi\sigma^2)-\frac{N}{2}\log|\Sigma_\tau|-\frac{1}{2\sigma^2}\left\{NT\hat{\sigma}^2+(\gamma-\hat{\gamma})^\top\mathbf{Z}^\top\Omega_\tau^{-1}\mathbf{Z}(\gamma-\hat{\gamma})\right\}$$

  where $\hat{\sigma}^2=(\mathbf{Y}-\mathbf{Z}\hat{\gamma})^\top\Omega_\tau^{-1}(\mathbf{Y}-\mathbf{Z}\hat{\gamma})/(NT)$

- Given any value of τ, $(\hat{\gamma},\hat{\sigma}^2)$ is the MLE of (γ,σ^2)

Maximum Likelihood Estimation of τ

- Useful identities:

$$\begin{aligned}\Sigma_\tau^{-1} &= \mathbf{I}_T - \frac{\tau}{1+\tau T}\mathbf{1}_T\mathbf{1}_T^\top = \mathbf{M}_{\mathbf{D}_i} + \frac{1}{T(1+\tau T)}\mathbf{1}_T\mathbf{1}_T^\top \\ \Omega_\tau^{-1} &= \mathbf{I}_N \otimes \Sigma_\tau^{-1} = \mathbf{M}_\mathbf{D} + \mathbf{I}_N \otimes \frac{g(\tau)}{T}\mathbf{1}_T\mathbf{1}_T^\top\end{aligned}$$

  where $\mathbf{M}_\mathbf{D} = \mathbf{I}_N \otimes \mathbf{M}_{\mathbf{D}_i}$ and $g(\tau) = 1/(1+\tau T)$
- Thus,

$$\mathbf{Z}^\top \Omega_\tau^{-1}\mathbf{Z} = \widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}} + \mathbf{Z}^\top\left(\mathbf{I}_N \otimes \frac{g(\tau)}{T}\mathbf{1}_T\mathbf{1}_T^\top\right)\mathbf{Z}$$

  where the second term equals

$$\frac{g(\tau)}{T}\sum_{i=1}^N Z_i^\top \mathbf{1}_T\mathbf{1}_T^\top Z_i = g(\tau) T\overline{\mathbf{Z}}^\top\overline{\mathbf{Z}}$$

  where $\overline{\mathbf{Z}}$ is a stacked matrix based on $\overline{\mathbf{Z}}_i = \frac{1}{T}Z_i^\top\mathbf{1}_T$

- Now the MLE of γ and σ^2 can be written as,

$$\begin{aligned}\hat{\gamma}_\tau &= (\widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{Z}} + g(\tau) T \overline{\mathbf{Z}}^\top \overline{\mathbf{Z}})^{-1} (\widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{Y}} + g(\tau) T \overline{\mathbf{Z}}^\top \bar{Y}) \\ \hat{\sigma}^2_\tau &= \frac{1}{NT} (\hat{\tilde{\epsilon}}^\top \hat{\tilde{\epsilon}} + g(\tau) T \hat{\bar{\epsilon}}^\top \hat{\bar{\epsilon}})\end{aligned}$$

where $\hat{\tilde{\epsilon}} = \widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}} \hat{\gamma}_\tau$ and $\hat{\bar{\epsilon}} = \bar{\mathbf{Y}} - \overline{\mathbf{Z}} \hat{\gamma}_\tau$

- Maximize the “concentrated” log-likelihood with respect to τ:

$$\hat{\tau} \;=\; \underset{\tau}{\operatorname{argmax}} \; -\frac{N}{2} \{ T \log \hat{\sigma}^2_\tau + \log(1 + \tau T) \}$$

where $|\Sigma_\tau| = 1 + \tau T$

Within-group and Between-group Interpretation

- Recall the within-group estimator: $\hat{\beta}_W = (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{Y}}$
- Between-group estimator: $\hat{\beta}_B = (\ddot{\mathbf{X}}^\top \ddot{\mathbf{X}})^{-1} \ddot{\mathbf{X}}^\top \ddot{\mathbf{Y}}$ where $\ddot{\mathbf{Y}}_i = \bar{y}_i - \bar{y}$, $\ddot{\mathbf{X}}_i = (\bar{x}_i - \bar{x})^\top$, and

$$\underbrace{\ddot{\mathbf{Y}}}_{NT \times 1} = \begin{pmatrix} \mathbf{1}_T(\bar{y}_1 - \bar{y}) \\ \vdots \\ \mathbf{1}_T(\bar{y}_N - \bar{y}) \end{pmatrix} \quad \text{and} \quad \underbrace{\ddot{\mathbf{X}}}_{NT \times K} = \begin{pmatrix} \mathbf{1}_T(\bar{x}_1 - \bar{x})^\top \\ \vdots \\ \mathbf{1}_T(\bar{x}_N - \bar{x})^\top \end{pmatrix}$$

- Random effects estimator as the weighted average:

$$\begin{aligned} \hat{\alpha}_{RE} &= \bar{y} - \hat{\beta}_{RE}^\top \bar{x} \\ \hat{\beta}_{RE} &= (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} + g(\tau) \ddot{\mathbf{X}}^\top \ddot{\mathbf{X}})^{-1} (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} \hat{\beta}_W + g(\tau) \ddot{\mathbf{X}}^\top \ddot{\mathbf{X}} \hat{\beta}_B) \end{aligned}$$

- $g(\tau) \to 0$ when $T \to \infty$ or $\tau \to \infty$
- Asymptotically, $\hat{\beta}_{RE} \sim \mathcal{N}(\beta,\ \sigma^2 (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} + g(\tau) \ddot{\mathbf{X}}^\top \ddot{\mathbf{X}})^{-1})$
- Under random effects model

$$\mathbb{V}(\hat{\beta}_{RE} \mid \mathbf{X}) \approx \sigma^2 (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} + g(\tau) \ddot{\mathbf{X}}^\top \ddot{\mathbf{X}})^{-1} \leq \mathbb{V}(\hat{\beta}_W \mid \mathbf{X}) \approx \sigma^2 (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1}$$

Shrinkage and BLUP

- Estimation of varying intercepts under the random effects model
- Empirical Bayes approach:
  - Bayes: likelihood + subjective prior
  - Empirical Bayes: likelihood + "objective" prior (estimated from data)
- The posterior of α_i

$$\mathcal{N}\left(\frac{\tau T}{1+\tau T}(\bar{y}_i - \beta^\top \bar{x}_i) + \frac{1}{1+\tau T}\alpha, \ \frac{\sigma^2\tau}{1+\tau T}\right)$$

- Plug in $\hat{\alpha}_{RE}, \hat{\beta}_{RE}, \hat{\tau}, \hat{\sigma}^2$ to obtain $\hat{\alpha}_i$ and its confidence interval
- $\hat{\alpha}_{i,RE}$ is the BLUP (Best Linear Unbiased Predictor) without the distributional assumption for α_i
- Shrinkage (partial pooling): weighted average of within-group mean and overall mean where the weight is a function of T and τ
- Borrowing strength: key idea for multilevel/hierarchical models, variable selection (ridge regression, LASSO, etc.)
- Bias-variance tradeoff

Fixed or Random Intercepts?

- Dilemma: Random effects impose additional assumptions but can be more efficient if the assumptions are correct
- Hausman specification test
  1. Test statistic

$$\begin{aligned} H &\equiv (\hat{\beta}_w - \hat{\beta}_{RE})^\top \mathbb{V}(\hat{\beta}_W - \hat{\beta}_{RE} \mid \mathbf{X})^{-1} (\hat{\beta}_w - \hat{\beta}_{RE}) \\ &= (\hat{\beta}_w - \hat{\beta}_{RE})^\top \{\mathbb{V}(\hat{\beta}_W \mid \mathbf{X}) - \mathbb{V}(\hat{\beta}_{RE} \mid \mathbf{X})\}^{-1} (\hat{\beta}_w - \hat{\beta}_{RE}) \end{aligned}$$

     Hausman shows that asymptotically $(\hat{\beta}_W - \hat{\beta}_{RE}) \perp\!\!\!\perp \hat{\beta}_{RE}$
  2. Null hypothesis: random effects model
  3. Asymptotic reference distribution: $H \sim \chi^2_K$
- Warning: the alternative hypothesis is that random effects model is wrong but fixed effects model is correct, but in practice both models could be wrong!

Hausman Test Proof

Lemma: Consider two consistent and asymptotically normal estimators of β and call them $\hat{\beta}_0$ and $\hat{\beta}_1$. Suppose $\hat{\beta}_0$ attains the Cramer-Rao lower bound. Then, $\mathrm{Cov}(\hat{\beta}_0, \hat{q})$ converges asymptotically to zero where $\hat{q} = \hat{\beta}_0 - \hat{\beta}_1$.

1. Define a new estimator $\hat{\beta}_2 = \hat{\beta}_0 + rA\hat{q}$ where r is a scalar and A is an arbitrary matrix to be chosen later
2. Show that $\hat{\beta}_2 \xrightarrow{p} \beta$ with the asymptotic variance
$$\mathbb{V}(\hat{\beta}_2) = \mathbb{V}(\hat{\beta}_0) + 2rA\mathrm{Cov}(\hat{q}, \hat{\beta}_0) + r^2 A\mathbb{V}(\hat{q})A^\top$$
3. Consider $F(r) = \mathbb{V}(\hat{\beta}_2) - \mathbb{V}(\hat{\beta}_0) \geq 0$
4. Choose $A = -\mathrm{Cov}(\hat{q}, \hat{\beta}_0)^\top$ and show that $F(r) < 0$ for a small value of r, which yields a contradiction unless $\mathrm{Cov}(\hat{q}, \hat{\beta}_0) = 0$
5. Finally, $\mathbb{V}(\hat{\beta}_1) = \mathbb{V}(\hat{q} + \hat{\beta}_0) = \mathbb{V}(\hat{q}) + \mathbb{V}(\hat{\beta}_0)$

An Example: Democratic Peace Debate

- *International Organization* special issue
- Green *et al.*, Oneal & Russett, Beck & Katz, King
- Dyadic analysis
- Effect of Democracy on bilateral trade (given here) and conflict (see later slide)
- Hausman test for pooled analysis vs. fixed effects

Variable[a]	*Pooled*	*Fixed effects*
GDP	1.182**	0.810**
	(0.008)	(0.015)
Population	−0.386**	0.752**
	(0.010)	(0.082)
Distance	−1.342**	Dropped: no within-group variation
	(0.018)	
Alliance	−0.745**	0.777**
	(0.042)	(0.136)
Democracy[b]	0.075**	−0.039**
	(0.002)	(0.003)
Lagged bilateral trade		
Constant	−17.331**	−47.994**
	(0.265)	(1.999)
	$N = 93{,}924$	$NT = 93{,}924$
		$N = 3{,}079$
		$T \geq 20$
Adjusted R^2	0.36	0.63

A Generalization of Random Effects Model

- Random effects model assumes $\alpha_i \perp\!\!\!\perp \mathbf{X}_i$
- "Correlated" random effects:

$$\alpha_i \mid \mathbf{X}_i \stackrel{\text{indep.}}{\sim} \mathcal{N}(\alpha + \xi^\top \bar{x}_i,\ \omega^2)$$

- Then, $\beta^\top x_{it} + \xi^\top \bar{x}_i = \beta^\top (x_{it} - \bar{x}_i) + (\beta + \xi)^\top \bar{x}_i$ implies

$$\mathbf{Z}_i = \begin{pmatrix} 1 & (x_{i1} - \bar{x}_i)^\top & \bar{x}_i^\top \\ \vdots & \vdots & \vdots \\ 1 & (x_{iT} - \bar{x}_i)^\top & \bar{x}_i^\top \end{pmatrix} \quad \text{and} \quad \gamma = \begin{pmatrix} \alpha \\ \beta \\ \lambda \end{pmatrix}$$

  where $\lambda = \beta + \xi$
- This leads to the surprising result:

$$\hat{\beta}_{CRE} = \hat{\beta}_W \quad \text{and} \quad \hat{\lambda}_{CRE} = \hat{\beta}_B$$

- Empirical Bayes estimate of α_i:

$$\frac{\hat{\tau} T}{1 + \hat{\tau} T} \hat{\alpha}_{i,FE} + \frac{1}{1 + \hat{\tau} T} (\bar{y} - \hat{\beta}_B^\top \bar{x} + (\hat{\beta}_B - \hat{\beta}_W)^\top \bar{x}_i)$$

Models with Varying Slopes and Intercepts

- Random effects model can be made richer and more flexible
- Linear mixed effects models:

$$\begin{aligned} Y_i \mid \mathbf{X}, \mathbf{Z}, \zeta & \overset{\text{indep.}}{\sim} \mathcal{N}(X_i\beta + Z_i\zeta_i,\ \Sigma_i) \\ \zeta_i \mid \mathbf{X}, \mathbf{Z} & \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Omega) \end{aligned}$$

  where $\mathbf{Z}$ is typically a subset of $\mathbf{X}$
- Estimating ζ_i without partial pooling is unrealistic when the dimension of Z_i is large
- Useful if intercepts/slopes differ across units and are of interest
- Multilevel/hierarchical models are extensions of this basic model (see Gelman and Hill (2007) for many interesting examples)
- Reduced form:

$$Y_i \mid \mathbf{X}, \mathbf{Z} \overset{\text{indep.}}{\sim} \mathcal{N}(X_i\beta,\ \Lambda_i)$$

  where $\Lambda_i = Z_i \Omega Z_i^\top + \Sigma_i$

Estimation of Linear Mixed Effects Models

- With known variance, the GLS is the MLE:

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i^\top \Lambda_i^{-1} X_i\right)^{-1} \sum_{i=1}^{N} X_i^\top \Lambda_i^{-1} Y_i$$

- Empirical Bayes estimate for ζ_i:

$$\begin{aligned} \hat{\zeta}_i &= (Z_i^\top \Sigma_i^{-1} Z_i + \Omega^{-1})^{-1} Z_i^\top \Sigma_i^{-1} (Y_i - X_i \hat{\beta}) \\ &= \Omega Z_i^\top \Lambda_i^{-1} (Y_i - X_i \hat{\beta}) \end{aligned}$$

- For known variance, $\hat{\zeta}_i$ attains the smallest MSE
- Restricted Maximum Likelihood (REML) to estimate variance where β is integrated out from the likelihood over improper prior
- `lmer()` in the `lme4` package

- Possible to obtain the MLE of all parameters at once via the *EM* algorithm treating ζ_i as missing data
- Fully Bayesian approach via Gibbs sampling

Generalized Linear Mixed Effects Models (GLMM)

- Extension of Linear Mixed Effects Models
  1. Linear predictor: $\eta_{it} = \beta^\top x_{it} + \zeta_i^\top z_{it}$
  2. Link function: $g(\mu_{it}) = \eta_{it}$ where $\mu_i = \mathbb{E}(y_{it} \mid \mathbf{X}, \mathbf{Z}, \zeta_i)$
  3. Random components:
     - $y_{it} \mid \mathbf{X}, \mathbf{Z}, \zeta_i \overset{\text{indep.}}{\sim} f(y \mid \eta_{it})$ an exponential-family distribution
     - $\zeta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Omega)$
- All diagnostics etc. for GLM can be applied
- The likelihood function:

$$\prod_{i=1}^{N} \left[\int \prod_{t=1}^{T} f(y_{it} \mid \eta_{it}) h(\zeta_i) d\zeta_i \right]$$

- In most cases, no analytical solution to the integral exists

Estimation of GLMM

- Decomposition: $y_{it} = g^{-1}(\eta_{it}) + \epsilon_{it}$
- Taylor expansion around current estimates $(\beta^{(t)}, \zeta^{(t)})$:

$$Y_i^{(t)} \approx X_i\beta + Z_i\zeta_i + \epsilon_i^{(t)} \quad \text{where} \quad Y_i^{(t)} = (\widehat{V}_i^{(t)})^{-1}(Y_i - \hat{\mu}_i) + X_i\hat{\beta}^{(t)} + Z_i\hat{\zeta}_i^{(t)}$$

  where $\widehat{V}_i$ is the diagonal matrix whose diagonal element corresponds the variance function $b''(\hat{\mu}_{it})$
- Iterated Weighted Least Squares where each iteration involves the optimization problem under a linear mixed effects model
- The resulting estimator can be justified as the penalized quasi-likelihood estimator
- The approximation can be poor
- Alternative approximation: Gaussian quadrature (`glmer()`)
- *MCEM* and *MCMC*

An Example: Modeling Latent Social Networks

- Hoff and Ward (*Political Analysis*, 2004)
- Modeling bilateral trade using cross-section dyadic data
- Model (export from i to j)

$$y_{ij} = \beta^\top x_{ij} + \underbrace{a_i + b_j + \gamma_{ij} + z_i^\top z_j}_{\epsilon_{ij}}$$

  where a_i is the sender effect, b_j is the receiver effect, γ_{ij} is the dyadic effect, z_i is a vector in the latent network space
- Random effects specification
  1. $(a_i, b_i) \sim \mathcal{N}(0, \Sigma)$
  2. $\gamma_{ij} = \gamma_{ji} \sim \mathcal{N}(0, \Phi)$
  3. $z_i \sim \mathcal{N}(0, \sigma^2 I)$

Generalized Estimating Equations

- If you are not interested in varying intercepts/slopes themselves, you need not estimate ζ_i in GLMM
- Model $\mathbb{E}(Y_i \mid \mathbf{X}) = g^{-1}(X_i\beta)$ rather than $\mathbb{E}(Y_i \mid \mathbf{X}, \mathbf{Z}, \zeta_i)$ (where $\mathbf{Z}$ is a subset of $\mathbf{X}$ though this does not have to be the case)
- Advantage: no need to assume a particular covariance of Y_i

$$V_i \ = \ \mathbb{V}(Y_i \mid \mathbf{X}) \ = \ \phi A_i(\beta)^{1/2} R(\alpha) A_i(\beta)^{1/2}$$

  where R is the "working" correlation matrix, A_i is a diagonal matrix of variances, and ϕ is a dispersion parameter
- Generalized estimating equation (GEE):

$$U(\beta) \ = \ \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)^{\top} V_i^{-1}(Y_i - \mu_i) = 0$$

- A review article: Zorn (*AJPS*, 2001)

Properties of GEE

- Close connection to the Method of Moments
- Asymptotic properties hold even when $R(\alpha)$ is misspecified (consistency and normality with robust standard error)

$$\sqrt{N}(\hat{\beta}_N - \beta_0) \xrightarrow{D} \mathcal{N}\left(0, \mathbb{E}(D_i^\top V_i^{-1} D_i)^{-1} \mathbb{E}(D_i^\top V_i^{-1} \mathbb{V}(Y_i \mid \mathbf{X}) V_i^{-1} D_i) \mathbb{E}(D_i^\top V_i^{-1} D_i)^{-1}\right)$$

where D_i is $\partial \mu_i / \partial \beta$

- Advantages of GEE
  1. Only need to get the mean right! (most efficient when you get the variance structure correct)
  2. Unlike the independence model with robust standard error, you get consistency as well as asymptotic normality
  3. Possible efficiency gain by accounting for correlation in the data
- GEE does assume exogeneity: you need to get the mean correct

Working Correlation Matrix and Estimation of GEE

- Popular choices:
  1. Independence: $R_i(t, t') = 0$
  2. Exchangeability: $R_i(t, t') = \alpha$
  3. AR(1): $R_i(t, t') = \alpha^{|t-t'|}$
  4. Stationary m-dependence: $R_i(t, t') = \begin{cases} \alpha_{t,t'} & \text{if } |t-t'| \leq m \\ 0 & \text{otherwise} \end{cases}$
  5. Unstructured: $R_i(t, t') = \alpha_{t,t'}$

- Iterative estimation procedure
  1. Use $\beta^{(t)}$ to update D_i and A_i
  2. Estimate $\mathbb{V}(Y_i \mid \mathbf{X})^{(t)} = \sum_{i=1}^N (Y_i - \mu^{(t)})^\top (Y_i - \mu^{(t)})/N$
  3. Compute Pearson residuals $\hat{\epsilon}_{it}^P$ and estimate $\phi^{(t)} = \sum_{i=1}^N \sum_{t=1}^T (\hat{\epsilon}_{it}^P)^2/NT$ and $R(\hat{\alpha}^{(t)})$, which give $V_i^{(t)}$
  4. Update β using one iteration of Fisher-scoring algorithm

$$\beta^{(t+1)} = \beta^{(t)} - \left\{ \sum_{i=1}^n (D_i^{(t)})^\top (V_i^{(t)})^{-1} D_i^{(t)} \right\}^{-1} \left\{ \sum_{i=1}^n D_i^{(t)} (V_i^{(t)})^{-1} (Y_i - \mu^{(t)}) \right\}$$

Fixed Effects in GLM

- Model:
$$\mathbb{E}(y_{it} \mid \mathbf{X}) \; = \; g^{-1}(\beta^\top x_{it} + \alpha_i)$$
- Incidental parameter problem (Neyman and Scott):
    - # of parameters goes to infinity as N tends to infinity (for fixed T)
    - Asymptotic properties of MLE are no longer guaranteed to hold
    - Especially problematic for small T and large N
- In the linear case, everything is fine because the sampling distribution of $\hat{\beta}$ does not depend on α_i
- In the nonlinear case, this is not generally the case
- Canonical example: fixed effects logistic regression with $T = 2$ and $x_{it} = \text{time dummy}$
    - Model: $\Pr(y_{it} = 1 \mid \mathbf{X}) = \frac{\exp(\alpha_i + \beta x_{it})}{1 + \exp(\alpha_i + \beta x_{it})}$
    - $\hat{\alpha}_i = \begin{cases} \infty & \text{if } (y_{i1}, y_{i2}) = (1, 1) \\ -\infty & \text{if } (y_{i1}, y_{i2}) = (0, 0) \end{cases}$
    - $\hat{\beta} \xrightarrow{P} 2\beta$

Conditional Likelihood Inference

- Maximize conditional likelihood given a sufficient statistic for α_i
- $S(Y)$ is said to be a sufficient statistic for θ if the conditional distribution of Y given $S(Y)$ does not depend on θ
- Properties (Andersen): asymptotically consistent and normal, less efficient than MLE
- A simple example:

$$y_{it} = \beta^\top x_{it} + \alpha_i + \epsilon_{it} \quad \text{where} \quad \epsilon_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

- Sufficient statistic for α_i is $\sum_{t=1}^T y_{it}$
- Conditional likelihood function:

$$\prod_{i=1}^N f\left(y_{i1}, \ldots, y_{iT} \;\middle|\; x_{it}, \sum_{t=1}^T y_{it}\right)$$
$$\propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^N \left(\sum_{t=1}^T \left\{(y_{it} - \bar{y}_i) - \beta^\top (x_{it} - \bar{x}_i)\right\}^2\right)\right]$$

Logistic Regression with Fixed Effects

- A sufficient statistic for α_i is again $\sum_{t=1}^{T} y_{it}$
- A special case: $T = 2$
  - $w_i = \begin{cases} 0 & \text{if } (y_{i1}, y_{i2}) = (1, 0) \\ 1 & \text{if } (y_{i1}, y_{i2}) = (0, 1) \end{cases}$
  - $\Pr(w_i = 1 \mid y_{i1} + y_{i2} = 1) = \frac{\exp(\beta^\top (x_{i2} - x_{i1}))}{1 + \exp(\beta^\top (x_{i2} - x_{i1}))}$
  - Conditional likelihood:

$$\prod_{i=1}^{N} \text{logit}^{-1}(\beta^\top (x_{i2} - x_{i1}))^{w_i} \{1 - \text{logit}^{-1}(\beta^\top (x_{i2} - x_{i1}))\}^{1 - w_i}$$

- The general case:

$$\prod_{i=1}^{N} \frac{\exp(\beta^\top \sum_{t=1}^{T} x_{it} y_{it})}{\sum_{d \in B_i} \exp(\beta^\top \sum_{t=1}^{T} x_{it} d_t)}$$

where $B_i = \{d = (d_1, \ldots, d_T) \mid d_t = 0 \text{ or } 1, \text{ and } \sum_{t=1}^{T} d_t = \sum_{t=1}^{T} y_{it}\}$

Estimation and Conditional Likelihood

- Equivalent to the partial likelihood function of the stratified Cox model in discrete time
  - Single discrete time period (rather than continuous)
  - All observations with $y_{it} = 1$ are considered as failures at time 1
  - All observations with $y_{it} = 0$ are considered as censored at time 1
  - Units correspond to strata
- In R, you use the following syntax or `clogit()`:

```
coxph(Surv(time = rep(1, N*T), status = y) ~
      x + strata(units), method = "exact")
```

- The exact calculation can be difficult when the number of time periods is large; use approximation

Limitations of Conditional Likelihood Approach

- Loss of information: e.g., observations with $\sum_{t=1}^{T} y_{it} = 0$ or T
- Sufficient statistics are not easily found
  - It works for logit, multinomial logit, Weibull etc.
  - But it does not work for probit, etc.
- Cannot estimate the quantities of interest
  - Only β can be estimated
  - Predicted probabilities, risk difference etc. cannot be estimated
  - Risk ratio, odds ratio can be estimated

- An alternative approach: correlated random effects
  - Recall that in the linear case these two approaches give the same estimate of β
  - In the nonlinear case, this does not hold but random effects can still be correlated with covariates
  - All quantities of interest can be estimated
  - A special case of GLMM

Back to Dirty Pool

- Democratic Peace: Effect of Democracy on militarized disputes
- Hausman test for pooled analysis vs. fixed effects
- Conditional likelihood: Peaceful dyads are dropped
- What is the effect size?
- Oneal and Russett estimate positive effects using fixed effects model for the data from 1885 (instead of 1952)
- Heterogeneous effects?

Variable	*Pooled*	*Fixed effects*
Contiguity	3.042**	1.902**
	(0.092)	(0.336)
Capability ratio (log)	0.102**	0.387**
	(0.024)	(0.139)
Growth[a]	−0.017	−0.059**
	(0.011)	(0.012)
Alliance	−0.234*	−1.066*
	(0.097)	(0.426)
Democracy[a]	−0.057**	−0.003
	(0.007)	(0.015)
Bilateral trade/GDP[a]	−0.194*	−0.072
	(0.087)	(0.186)
Lagged dispute		
Constant	−5.809**	
	(0.090)	
N	93,755	93,755[b]
Log likelihood	−3,688.06	−1,546.53
χ^2	1,186.43	75.75
Degrees of freedom	6	6
Prob > χ^2	<0.0001	<0.0001

Modeling Dynamics

- Previous approaches: treating dynamics as a nuisance
- If dynamics are of substantive interest, they should be modeled
- Dynamic linear models (DLMs):

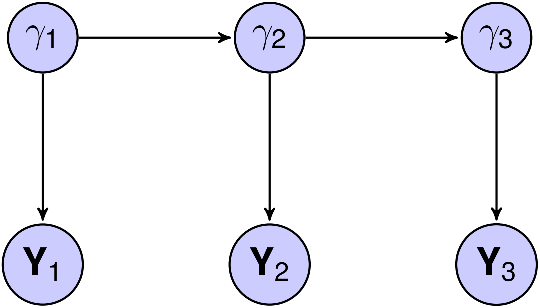
$$y_{it} \overset{\text{indep.}}{\sim} \mathcal{N}(X_{it}^\top \beta + Z_{it}^\top \gamma_t,\ \sigma^2)$$

  where time-varying coefficients γ_t has a random-walk prior,

$$\gamma_t \overset{\text{indep.}}{\sim} \mathcal{N}(\gamma_{t-1},\ \Sigma)$$

  for $t = 2, \ldots, T$ and $\gamma_1 \sim \mathcal{N}(\mu_0, \Omega_0)$
- Can add the second level covariates: $\gamma_t \overset{\text{indep.}}{\sim} \mathcal{N}(V_t \gamma_{t-1}, \Sigma)$
- A special case of state-space models and Markov models

Dependence through the Simple Markov Structure



Example: Ideal Point Models

- Ideal point model (Clinton, Jackman & Rivers, 2004; aka Item Response Theory):

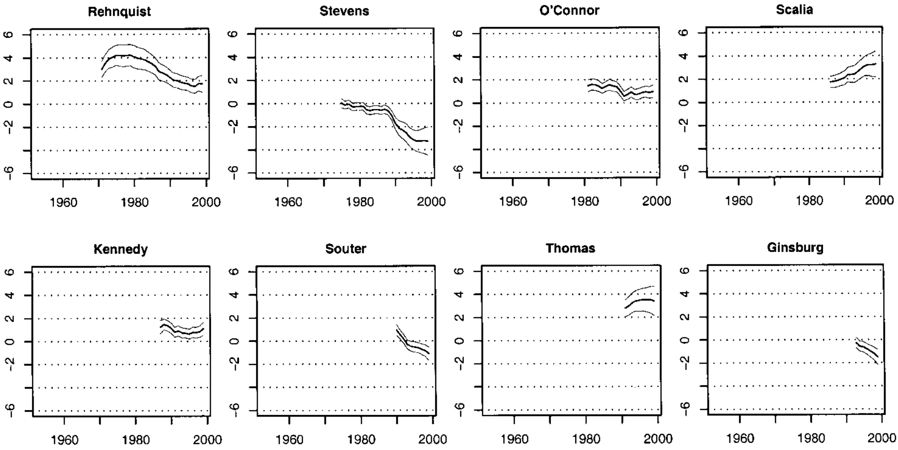
$$y_{ij}^* \quad = \quad \alpha_j + \beta_j x_i + \epsilon_{ij}$$

  where $y_{ij}^* \geq 0$ if $y_{ij} = 1$ ("yea") and $y_{ij}^* < 0$ if $y_{ij} = 0$ ("nay")
  - α_j: item difficulty
  - β_j: item discrimination
  - x_i: ideological position, i.e., ideal point
  - ϵ_i: spatial voting model error, typically assumed to be $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$
- Dynamic ideal point model (Martin and Quinn, 2002):

$$\begin{aligned} y_{ijt}^* &= \alpha_{jt} + \beta_{jt} x_{it} + \epsilon_{ijt} \\ x_{it} &= x_{i,t-1} + \eta_{it} \end{aligned}$$

  where $\eta_{it} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \omega_i^2)$
- What's the key identification assumption for the dynamic model?

Estimated Ideal Points for Supreme Court Justices



EM Algorithm for DLM

- Complete-data likelihood function:

$$\begin{aligned} & p(\mathbf{Y}, \gamma \mid \mathbf{X}, \mathbf{Z}; \beta, \sigma^2, \Sigma, \mu_0, \Omega_0) \\ = \;& p(\gamma_1; \mu_0, \Omega_0) \prod_{t=2}^{T} p(\gamma_t \mid \gamma_{t-1}; \Sigma) \prod_{t=1}^{T} p(\mathbf{Y}_t \mid \mathbf{X}_t, \mathbf{Z}_t, \gamma_t; \beta, \sigma^2) \end{aligned}$$

where $\mathbf{Y}_t \overset{\text{indep.}}{\sim} \mathcal{N}(\mathbf{X}_t \beta + \mathbf{Z}_t \gamma_t,\ \sigma^2 \mathbf{I}_N)$

- EM updates:
  1. $\mu_0 = \mathbb{E}(\gamma_1)$ and $\Omega_0 = \mathbb{E}(\gamma_1 \gamma_1^\top) - \mathbb{E}(\gamma_1)\mathbb{E}(\gamma_1)^\top$
  2. $\Sigma = \frac{1}{T-1} \sum_{t=2}^{T} \{\mathbb{E}(\gamma_t \gamma_t^\top) - \mathbb{E}(\gamma_t \gamma_{t-1}^\top) - \mathbb{E}(\gamma_{t-1} \gamma_t^\top) + \mathbb{E}(\gamma_{t-1} \gamma_{t-1}^\top)\}$
  3. $\beta = (\sum_{t=1}^{T} \mathbf{X}_t^\top \mathbf{X}_t)^{-1} \sum_{t=1}^{T} \mathbf{X}_t^\top \{\mathbf{Y}_t - \mathbf{Z}_t \mathbb{E}(\gamma_t)\}$
  4. $\sigma^2 = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} \{\tilde{y}_{it}^2 - 2\tilde{y}_{it} Z_{it}^\top \mathbb{E}(\gamma_t) + Z_{it}^\top \mathbb{E}(\gamma_t \gamma_t^\top) Z_{it}\}$
     where $\tilde{y}_{it} = y_{it} - X_{it}^\top \beta$

Computation for E-step

- Must compute $\mathbb{E}(\gamma_t)$, $\mathbb{E}(\gamma_t \gamma_t^\top)$ and $\mathbb{E}(\gamma_t \gamma_{t-1}^\top)$ conditional on *all* data
- Define

$$\begin{aligned} \alpha(\gamma_t) &= p(\gamma_t \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_t) \\ \delta(\gamma_t) &= p(\mathbf{Y}_{t+1}, \ldots, \mathbf{Y}_T \mid \gamma_t) \end{aligned}$$

- The posterior is given by,

$$p(\gamma_t \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T) = c_t \, \alpha(\gamma_t) \, \delta(\gamma_t)$$

  where $c_t = 1/p(\mathbf{Y}_{t+1}, \ldots, \mathbf{Y}_T \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_t)$ is a normalizing constant
- The joint distribution of $(\mathbf{Y}_1, \ldots, \mathbf{Y}_T)$ and $(\gamma_1, \ldots, \gamma_T)$ is Gaussian
- $\alpha(\gamma_t)$, $\delta(\gamma_t)$, and $\alpha(\gamma_t)\delta(\gamma_t)$ are all Gaussian
- Forward-backward algorithm thorough Kalman filtering

Forward Recursion

$$
\begin{aligned}
\alpha(\gamma_t) &= \int p(\gamma_t, \gamma_{t-1} \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_t)\, d\gamma_{t-1} \\
&= \int \frac{p(\gamma_t, \gamma_{t-1}, \mathbf{Y}_t \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_{t-1})}{p(\mathbf{Y}_t \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_{t-1})}\, d\gamma_{t-1} \\
&\propto \int p(\mathbf{Y}_t, \gamma_t \mid \gamma_{t-1}, \mathbf{Y}_1, \ldots, \mathbf{Y}_{t-1})\, p(\gamma_{t-1} \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_{t-1})\, d\gamma_{t-1} \\
&\propto p(\mathbf{Y}_t \mid \gamma_t) \int p(\gamma_t \mid \gamma_{t-1})\, \alpha(\gamma_{t-1})\, d\gamma_{t-1}
\end{aligned}
$$

Thus, $\alpha(\gamma_t) = \mathcal{N}(\mu_t, \Omega_t)$ is obtained as,

$$
\begin{aligned}
& \phi(\gamma_t; \mu_t, \Omega_t) \\
\propto\ & \phi(\mathbf{Y}_t; \mathbf{X}_t\beta + \mathbf{Z}_t\gamma_t, \sigma^2\mathbf{I}) \int \phi(\gamma_t; \gamma_{t-1}, \Sigma)\phi(\gamma_{t-1}; \mu_{t-1}, \Omega_{t-1})\, d\gamma_{t-1} \\
\propto\ & \phi(\mathbf{Y}_t; \mathbf{X}_t\beta + \mathbf{Z}_t\gamma_t, \sigma^2\mathbf{I})\, \phi(\gamma_t; \mu_{t-1}, \mathbf{Q}_{t-1})
\end{aligned}
$$

where $\mathbf{Q}_{t-1} = \Omega_{t-1} + \Sigma$.

Finally, Bayes rule gives,

$$\begin{aligned}\mu_t &= \Omega_t(\mathbf{Q}_{t-1}^{-1}\mu_{t-1} + \sigma^{-2}\mathbf{Z}_t^\top \widetilde{\mathbf{Y}}_t) \\ \Omega_t &= (\mathbf{Q}_{t-1}^{-1} + \sigma^{-2}\mathbf{Z}_t^\top \mathbf{Z}_t)^{-1}\end{aligned}$$

We can further simplify using the Woodbury formula,

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

which implies

$$\begin{aligned}\Omega_t &= \mathbf{Q}_{t-1} - \mathbf{Q}_{t-1}\mathbf{Z}_t^\top(\sigma^2\mathbf{I} + \mathbf{Z}_t\mathbf{Q}_{t-1}\mathbf{Z}_t^\top)^{-1}\mathbf{Z}_t\mathbf{Q}_{t-1} \\ &= (\mathbf{I} - \mathbf{R}_t\mathbf{Z}_t)\mathbf{Q}_{t-1}\end{aligned}$$

where $\mathbf{R}_t = \mathbf{Q}_{t-1}\mathbf{Z}_t^\top(\sigma^2\mathbf{I} + \mathbf{Z}_t\mathbf{Q}_{t-1}\mathbf{Z}_t^\top)^{-1}$. Finally, note another rule:

$$(\mathbf{A}^{-1} + \mathbf{B}^\top\mathbf{C}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{C}^{-1} = \mathbf{A}\mathbf{B}^\top(\mathbf{B}\mathbf{A}\mathbf{B}^\top + \mathbf{C})^{-1}$$

Then, we have

$$\begin{aligned}\Omega_t\mathbf{Z}_t^\top\sigma^{-2}\mathbf{I} &= \mathbf{Q}_{t-1}\mathbf{Z}_t^\top(\mathbf{Z}_t\mathbf{Q}_{t-1}\mathbf{Z}_t^\top + \sigma^2\mathbf{I})^{-1} = \mathbf{R}_t \\ \mu_t &= \mu_{t-1} + \mathbf{R}_t(\widetilde{\mathbf{Y}}_t - \mathbf{Z}_t\mu_{t-1})\end{aligned}$$

Backward Recursion

$$
\begin{aligned}
p(\gamma_t \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T) &= \int p(\gamma_t, \gamma_{t+1} \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T)\, d\gamma_{t+1} \\
&= \int p(\gamma_t \mid \gamma_{t+1}, \mathbf{Y}_1, \ldots, \mathbf{Y}_t) p(\gamma_{t+1} \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T)\, d\gamma_{t+1}
\end{aligned}
$$

From the forward recursion, we have,

$$
\begin{pmatrix} \gamma_{t+1} \\ \gamma_t \end{pmatrix} \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_t \sim \mathcal{N}\left(\begin{pmatrix} \mu_t \\ \mu_t \end{pmatrix}, \begin{pmatrix} \Gamma_t + \omega_x^2 & \Gamma_t \\ \Gamma_t & \Gamma_t \end{pmatrix} \right)
$$

Then, the conditional distribution is given by,

$$
\begin{aligned}
& \gamma_t \mid \gamma_{t+1}, \mathbf{Y}_1, \ldots, \mathbf{Y}_T \\
\sim\ & \mathcal{N}(\mu_t + \Gamma_t(\Gamma_t + \omega_x^2)^{-1}(\gamma_{t+1} - \mu_t),\ \Gamma_t - \Gamma_t(\Gamma_t + \omega_x^2)^{-1}\Gamma_t)
\end{aligned}
$$

Let's assume $p(\gamma_{t+1} \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T) = \mathcal{N}(m_{t+1}, S_{t+1})$. Then, we have,

$$\begin{aligned} m_t &= \mathbb{E}(\gamma_t \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T) \\ &= \mathbb{E}\{\mathbb{E}(\gamma_t \mid \gamma_{t+1}, \mathbf{Y}_1, \ldots, \mathbf{Y}_t) \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T\} \\ &= \mu_t + \Gamma_t(\Gamma_t + \omega_x^2)^{-1}(m_{t+1} - \mu_t) \end{aligned}$$

Finally,

$$\begin{aligned} S_t &= \mathbb{V}(\gamma_t \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T) \\ &= \mathbb{V}\{\mathbb{E}(\gamma_t \mid \gamma_{t+1}, \mathbf{Y}_1, \ldots, \mathbf{Y}_t) \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T\} \\ &\quad + \mathbb{E}\{\mathbb{V}(\gamma_t \mid \gamma_{t+1}, \mathbf{Y}_1, \ldots, \mathbf{Y}_t) \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T\} \\ &= \mathbb{V}\{\mu_t + \Gamma_t(\Gamma_t + \omega_x^2)^{-1}(\gamma_{t+1} - \mu_t) \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T\} \\ &\quad + \Gamma_t - \Gamma_t(\Gamma_t + \omega_x^2)^{-1}\Gamma_t \\ &= \Gamma_t + \Gamma_t(\Gamma_t + \omega_x^2)^{-1}\{S_{t+1} - (\Gamma_t + \omega_x^2)\}(\Gamma_t + \omega_x^2)^{-1}\Gamma_t \end{aligned}$$

Concluding Remarks

- Longitudinal data: opportunities to model within-unit and across-time variation as well as across-unit variation
- Old debates: fixed or random intercepts
- GLMM: a generalization of random effects models
- Model slopes as well as intercepts: importance of substantive theory
- Structural modeling and causal inference
- For causal inference, getting the conditional mean right is essential
- Static models ignore the systematic dependence on past outcomes: all dependence is in the error term