

Applied Regression Models for Longitudinal Data

Kosuke Imai

Princeton University

November 10, 2010

Readings

- (Required) Hayashi, *Econometrics*, Chapter 5
- (Required) “Dirty Pool” papers referenced in the slides
- (Suggested) Wooldrich, *Econometric Analysis of Cross Section and Panel Data*, Chapter 10 and relevant sections of Part IV
- (Suggested) Gelman and Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Part 2A, Cambridge University Press.
- (Reference) Hsiao, *Analysis of Panel Data*, 2nd ed., Cambridge University Press.
- (Reference) Diggle, *et al. Analysis of Longitudinal Data* 2nd ed., Oxford 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 $\alpha_i = \alpha + \delta^\top u_i$

- **Exogeneity** given x_{it} and u_i :

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

- 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)

$$\mathbb{V}(\epsilon_j | \mathbf{X}) = \sigma^2 I_T$$

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

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

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

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 7 and 11)
- Projection of \mathbf{Y} onto $\mathcal{S}^\perp(\mathbf{D})$
- Thus, $\tilde{\mathbf{Y}} = \mathbf{M}_D \mathbf{Y}$ where $\mathbf{M}_D = \mathbf{I}_{NT} - \mathbf{D}(\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top$
- Also, $\tilde{\mathbf{X}} = \mathbf{M}_D \mathbf{X}$
- This is a partitioned regression!
- Then, $\hat{\beta}_{FE} = \hat{\beta}_W$ (see Question 1 of POL572 Problem Set 5 ☺)
- Also, $\hat{\epsilon} = \mathbf{Y} - \mathbf{Z}\hat{\gamma}_{FE} = \tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\beta}_W$
- The lower-right block of $(\mathbf{Z}^\top \mathbf{Z})^{-1}$ equals $(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}$
- Thus, $\hat{\beta}_W | \mathbf{X} \sim \mathcal{N}(\beta, \sigma^2(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1})$

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):

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

where the “meat” is estimated by $\sum_{i=1}^N \tilde{\mathbf{X}}_i^\top \hat{\epsilon}_i \hat{\epsilon}_i^\top \tilde{\mathbf{X}}_i$

- 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

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

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

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

- Could relax the invariance assumption: $\hat{\rho}_t = \mathbb{E}(\widehat{\tilde{\epsilon}_t\tilde{\epsilon}_t} | \mathbf{X}) = \hat{\epsilon}_t \hat{\epsilon}_t^\top$
- These robust standard errors can be applied to any linear models (with or without fixed effects)

Random Effects Model

- Dimension reduction is desirable when N is large relative to T
- “Random” intercepts: a prior distribution on α_j

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

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

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

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

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

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

Estimation of Random Effects Model

- Log-likelihood function:

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

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

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

$$|\Sigma_\tau| = 1 + \tau T \quad \text{and} \quad \Sigma_\tau^{-1} = \mathbf{I}_T - \frac{\tau}{1 + \tau T} \mathbf{1}_T \mathbf{1}_T^\top$$

- Then, using $g(\tau) = 1/(1 + \tau T)$, we have

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

where the i th row of $\bar{\mathbf{Z}}$ is $[1 \ \bar{x}_i^\top]$ and $\hat{\bar{\epsilon}} = \bar{y} - \bar{\mathbf{Z}}\hat{\gamma}$

- Concentrated log-likelihood: $-\frac{N}{2} \{T \log 2\pi \hat{\sigma}^2 + \log(1 + \tau T) + T\}$

Within-group and Between-group Interpretation

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

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

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

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

Shrinkage and BLUP

- Estimation of varying intercepts under the random effects model
- Empirical Bayes approach (Details in POL574)
 - Bayes: likelihood + subjective prior
 - Empirical Bayes: likelihood + “objective” prior (estimated from data)
- The posterior of α_j

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

- 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_j
- 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

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

Hausman shows that asymptotically $(\hat{\beta}_w - \hat{\beta}_{RE}) \perp \hat{\beta}_{RE}$

- 2 Null hypothesis: random effects model
 - 3 Asymptotic reference distribution: $H \sim \chi_K^2$
- 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!

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** (0.008)	0.810** (0.015)
Population	-0.386** (0.010)	0.752** (0.082)
Distance	-1.342** (0.018)	Dropped: no within-group variation
Alliance	-0.745** (0.042)	0.777** (0.136)
Democracy ^b	0.075** (0.002)	-0.039** (0.003)
Lagged bilateral trade		
Constant	-17.331** (0.265)	-47.994** (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 $\alpha_j \perp \mathbf{X}_j$
- “Correlated” random effects:

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

- Then, $\beta^\top \mathbf{x}_{it} + \xi^\top \bar{\mathbf{x}}_j = \beta^\top (\mathbf{x}_{it} - \bar{\mathbf{x}}_j) + (\beta + \xi)^\top \bar{\mathbf{x}}_j$ implies

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

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_j :

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

Models with Varying Slopes and Intercepts

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

$$Y_i | \mathbf{X}, \mathbf{Z}, \zeta \stackrel{\text{indep.}}{\sim} \mathcal{N}(\mathbf{X}_i\beta + \mathbf{Z}_i\zeta_i, \Sigma_i)$$

$$\zeta_i | \mathbf{X}, \mathbf{Z} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Omega)$$

where \mathbf{Z} is typically a subset of \mathbf{X}

- Estimating ζ_i without partial pooling is unrealistic when the dimension of \mathbf{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:

$$Y_i | \mathbf{X}, \mathbf{Z} \stackrel{\text{indep.}}{\sim} \mathcal{N}(\mathbf{X}_i\beta, \Lambda_i)$$

where $\Lambda_i = \mathbf{Z}_i\Omega\mathbf{Z}_i^\top + \Sigma_i$

Estimation of Linear Mixed Effects Models

- With known variance, the GLS is the MLE:

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

- Empirical Bayes estimate for ζ_i :

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

- 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: $\eta_{it} = \beta^\top \mathbf{x}_{it} + \zeta_i^\top \mathbf{z}_{it}$
 - 2 Link function: $g(\mu_{it}) = \eta_{it}$ where $\mu_{it} = \mathbb{E}(y_{it} | \mathbf{X}, \mathbf{Z}, \zeta_i)$
 - 3 Random components:
 - $y_{it} | \mathbf{X}, \mathbf{Z}, \zeta_i \stackrel{\text{indep.}}{\sim} f(y | \eta_{it})$ an exponential-family distribution
 - $\zeta_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Omega)$
- All diagnostics etc. for GLM can be applied
- The likelihood function:

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

- 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 $(\beta^{(t)}, \zeta^{(t)})$:

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

where $\widehat{\mathbf{V}}_i$ 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 $\mathbb{E}(Y_i | \mathbf{X}) = g^{-1}(X_i \beta)$ rather than $\mathbb{E}(Y_i | \mathbf{X}, \mathbf{Z}, \zeta_i)$ (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

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

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)

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

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 $\nabla(Y_i | \mathbf{X})^{(t)} = \sum_{i=1}^N (Y_i - \mu^{(t)})^\top (Y_i - \mu^{(t)}) / N$
 - 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:

$$\mathbb{E}(y_{it} | \mathbf{X}) = g^{-1}(\beta^\top x_{it} + \alpha_j)$$

- 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_j
- 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: $\Pr(y_{it} = 1 | \mathbf{X}) = \frac{\exp(\alpha_j + \beta x_{it})}{1 + \exp(\alpha_j + \beta x_{it})}$
 - $\hat{\alpha}_j = \begin{cases} \infty & \text{if } (y_{i1}, y_{i2}) = (1, 1) \\ -\infty & \text{if } (y_{i1}, y_{i2}) = (0, 0) \end{cases}$
 - $\hat{\beta} \xrightarrow{P} 2\beta$

Conditional Likelihood Inference

- Maximize conditional likelihood given a sufficient statistic for α_j
- $S(Y)$ 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:

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

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

$$\prod_{i=1}^N f\left(y_{i1}, \dots, y_{iT} \mid x_{it}, \sum_{t=1}^T y_{it}\right)$$

$$\propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^N \left\{ \sum_{t=1}^T (y_{it}^2 - \bar{y}_i^2) - \beta^\top (x_{it} - \bar{x}_i)^2 \right\}\right]$$

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

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?

<i>Variable</i>	<i>Pooled</i>	<i>Fixed effects</i>
Contiguity	3.042** (0.092)	1.902** (0.336)
Capability ratio (log)	0.102** (0.024)	0.387** (0.139)
Growth ^a	-0.017 (0.011)	-0.059** (0.012)
Alliance	-0.234* (0.097)	-1.066* (0.426)
Democracy ^a	-0.057** (0.007)	-0.003 (0.015)
Bilateral trade/GDP ^a	-0.194* (0.087)	-0.072 (0.186)
Lagged dispute		
Constant	-5.809** (0.090)	
<i>N</i>	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

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