

GSERM - St. Gallen 2024

Analyzing Panel Data

June 14, 2024

Start with:

$$Y_i^* = \mathbf{X}_i\beta + u_i$$

$$Y_i = 0 \text{ if } Y_i^* < 0$$

$$Y_i = 1 \text{ if } Y_i^* \geq 0$$

So:

$$\begin{aligned}\Pr(Y_i = 1) &= \Pr(Y_i^* \geq 0) \\ &= \Pr(\mathbf{X}_i\beta + u_i \geq 0) \\ &= \Pr(u_i \geq -\mathbf{X}_i\beta) \\ &= \Pr(u_i \leq \mathbf{X}_i\beta) \\ &= \int_{-\infty}^{\mathbf{X}_i\beta} f(u) du\end{aligned}$$

“Standard logistic” PDF:

$$\Pr(u) \equiv \lambda(u) = \frac{\exp(u)}{[1 + \exp(u)]^2}$$

CDF:

$$\begin{aligned}\Lambda(u) &= \int \lambda(u) du \\ &= \frac{\exp(u)}{1 + \exp(u)} \\ &= \frac{1}{1 + \exp(-u)}\end{aligned}$$

Logistic \rightarrow “Logit”

$$\begin{aligned}\Pr(Y_i = 1) &= \Pr(Y_i^* > 0) \\ &= \Pr(u_i \leq \mathbf{X}_i\boldsymbol{\beta}) \\ &= \Lambda(\mathbf{X}_i\boldsymbol{\beta}) \\ &= \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}\end{aligned}$$

$$\text{(equivalently)} = \frac{1}{1 + \exp(-\mathbf{X}_i\boldsymbol{\beta})}$$

$$L_i = \left(\frac{\exp(\mathbf{X}_i\beta)}{1 + \exp(\mathbf{X}_i\beta)} \right)^{Y_i} \left[1 - \left(\frac{\exp(\mathbf{X}_i\beta)}{1 + \exp(\mathbf{X}_i\beta)} \right) \right]^{1-Y_i}$$

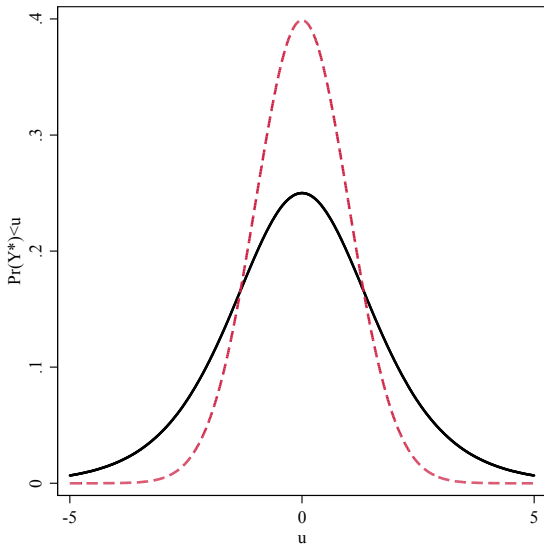
$$L = \prod_{i=1}^N \left(\frac{\exp(\mathbf{X}_i\beta)}{1 + \exp(\mathbf{X}_i\beta)} \right)^{Y_i} \left[1 - \left(\frac{\exp(\mathbf{X}_i\beta)}{1 + \exp(\mathbf{X}_i\beta)} \right) \right]^{1-Y_i}$$

$$\begin{aligned} \ln L &= \sum_{i=1}^N Y_i \ln \left(\frac{\exp(\mathbf{X}_i\beta)}{1 + \exp(\mathbf{X}_i\beta)} \right) + \\ &\quad (1 - Y_i) \ln \left[1 - \left(\frac{\exp(\mathbf{X}_i\beta)}{1 + \exp(\mathbf{X}_i\beta)} \right) \right] \end{aligned}$$

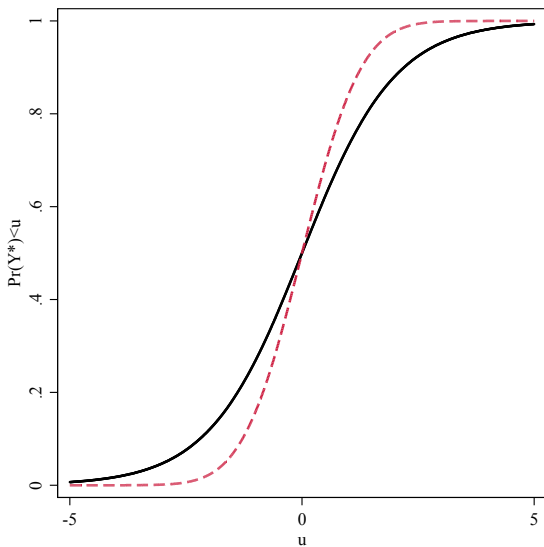
$$\Pr(u) \equiv \phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

$$\Phi(u) = \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$$

Standard Normal and Logistic PDFs



Standard Normal and Logistic CDFs



$$\begin{aligned}\Pr(Y_i = 1) &= \Phi(\mathbf{X}_i\beta) \\ &= \int_{-\infty}^{\mathbf{X}_i\beta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\mathbf{X}_i\beta)^2}{2}\right) d\mathbf{X}_i\beta\end{aligned}$$

$$L = \prod_{i=1}^N [\Phi(\mathbf{X}_i\beta)]^{Y_i} [1 - \Phi(\mathbf{X}_i\beta)]^{(1-Y_i)}$$

$$\ln L = \sum_{i=1}^N Y_i \ln \Phi(\mathbf{X}_i\beta) + (1 - Y_i) \ln [1 - \Phi(\mathbf{X}_i\beta)]$$

One-way unit effects:

$$Y_{it} = f(\mathbf{X}_{it}\beta + \alpha_i + u_{it})$$

So, think about logit first:

$$\Pr(Y_{it} = 1) = \frac{\exp(\mathbf{X}_{it}\beta + \alpha_i)}{1 + \exp(\mathbf{X}_{it}\beta + \alpha_i)} \equiv \Lambda(\mathbf{X}_{it}\beta + \alpha_i)$$

Incidental Parameters:

- Nonlinearity \rightarrow inconsistency in both $\hat{\alpha}$ s and $\hat{\beta}$.
- Anderson's *unconditional* estimator:

$$L^U = \prod_{i=1}^N \prod_{t=1}^T \Lambda(\mathbf{X}_{it} + \alpha_i)^{Y_{it}} [1 - \Lambda(\mathbf{X}_{it} + \alpha_i)]^{1-Y_{it}}$$

- Chamberlain's *conditional* estimator:

$$L^C = \prod_{i=1}^N \Pr \left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots, Y_{iT} = y_{iT} \mid \sum_{t=1}^T Y_{it} \right)$$

Fixed-Effects (continued)

Intuition: Suppose we have $T = 2$. That means that:

- $\Pr(Y_{i1} = 0 \text{ and } Y_{i2} = 0 \mid \sum_T Y_{it} = 0) = 1.0$
- $\Pr(Y_{i1} = 1 \text{ and } Y_{i2} = 1 \mid \sum_T Y_{it} = 2) = 1.0$

and:

$$\Pr\left(Y_{i1} = 0 \text{ and } Y_{i2} = 1 \mid \sum_T Y_{it} = 1\right) = \frac{\Pr(0, 1)}{\Pr(0, 1) + \Pr(1, 0)}$$

with a similar statement for $\Pr(Y_{i1} = 1 \text{ and } Y_{i2} = 0 \mid \sum_T Y_{it} = 1)$.

The Point:

$\sum_{t=1}^T Y_{it}$ is a sufficient statistic for α_i , so conditioning on it \equiv “fixed effects.”

Things to bear in mind:

- Fixed effects = no estimates for β_b
- Interpretation: per logit, but $|\hat{\alpha}_j$.
- Everything above is for **logit**...
 - For FE probit, there is no conditional model
 - Unconditional / “brute force” FE probit is biased (see [here](#) and [here](#))
- BTSCS in international relations: [Green et al. \(2001\)](#) vs. [Beck & Katz \(2001\)](#) (“Dirty Pool” debate)

Model is:

$$\begin{aligned} Y_{it}^* &= \mathbf{X}_{it}\beta + u_{it} \\ Y_{it} &= 0 \text{ if } Y_{it}^* \leq 0 ; \\ &= 1 \text{ if } Y_{it}^* > 0 \end{aligned}$$

with:

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

with $\eta_{it} \sim \text{i.i.d. } N(0,1)$, and $\alpha_i \sim N(0, \sigma_\alpha^2)$. This implies:

$$\text{Var}(u_{it}) = 1 + \sigma_\alpha^2$$

and so:

$$\text{Corr}(u_{it}, u_{is}, t \neq s) \equiv \rho = \frac{\sigma_\alpha^2}{1 + \sigma_\alpha^2}$$

which means that we can write $\sigma_\alpha^2 = \left(\frac{\rho}{1-\rho} \right)$.

Probit:

$$\begin{aligned} L_i &= \text{Prob}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots Y_{iT} = y_{iT}) \\ &= \int_{-\infty}^{X_{i1}\beta} \int_{-\infty}^{X_{i2}\beta} \dots \int_{-\infty}^{X_{iT}\beta} \phi(u_{i1}, u_{i2} \dots u_{iT}) du_{iT} \dots du_{i2} du_{i1} \end{aligned}$$

Logit:

$$\begin{aligned} L_i &= \text{Prob}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots Y_{iT} = y_{iT}) \\ &= \int_{-\infty}^{X_{i1}\beta} \int_{-\infty}^{X_{i2}\beta} \dots \int_{-\infty}^{X_{iT}\beta} \lambda(u_{i1}, u_{i2} \dots u_{iT}) du_{iT} \dots du_{i2} du_{i1} \end{aligned}$$

Solution?

$$\phi(u_{i1}, u_{i2}, \dots u_{iT}) = \int_{-\infty}^{\infty} \phi(u_{i1}, u_{i2}, \dots u_{iT} \mid \alpha_i) \phi(\alpha_i) d\alpha_i$$

- $\hat{\rho}$ = proportion of the variance due to the α_i s.
- Implementation: Gauss-Hermite quadrature or MCMC.
- Best with N large and T small.
- Critically requires $\text{Cov}(\mathbf{X}, \alpha) = 0$ (see notes re: Chamberlain's CRE Estimator).

Unit Effects in Practice - Some Simulations

Start with:

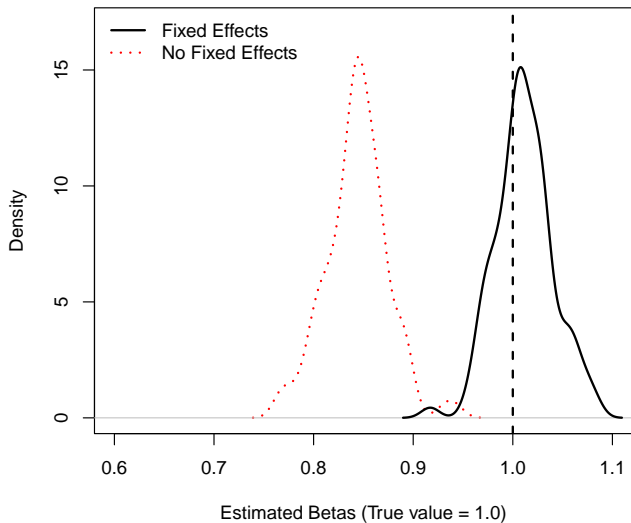
$$\begin{aligned} Y_{it}^* &= 0 + (1 \times X_{it}) + (1 \times D_{it}) + (1 \times \alpha_i) + u_{it} \\ Y_{it} \in \{0, 1\} &= f(Y_{it}^*) \end{aligned}$$

where:

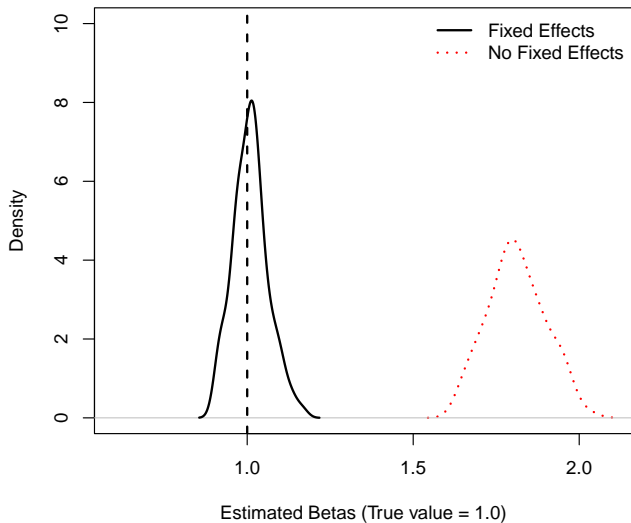
- $\alpha_i \sim N(0, 1)$
- $X_{it} \sim N(0, \sigma_X^2)$
- $D_{it} \in \{0, 1\}$
- $\text{Cov}(X_{it}, \alpha_i) = \{0, 0.69\}$
- $\text{Cov}(D_{it}, \alpha_i) = 0$
- $f(\cdot) = \{\text{logit}, \text{probit}\}$ (as appropriate)

and $N = T = 100$.

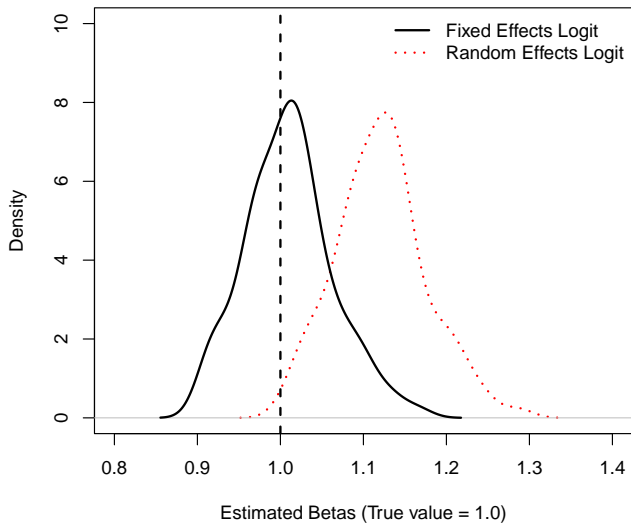
Logit $\hat{\beta}_X$ s for $\text{Cov}(X_{it}, \alpha_i) = 0$



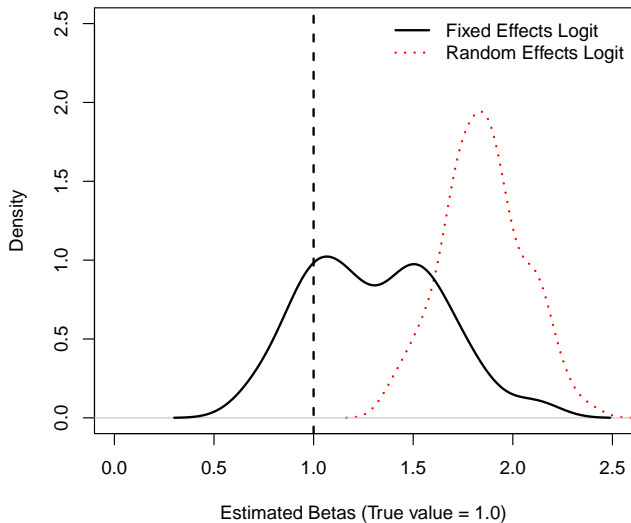
Logit $\hat{\beta}_{Xs}$ for $\text{Cov}(X_{it}, \alpha_i) \approx 0.69$



Logit $\hat{\beta}_{Xs}$ for $\text{Cov}(X_{it}, \alpha_i) \approx 0.69$



Same Plot, but with $T = 5...$



R

- `pglm` (panel GLMs) (maximum likelihood + quadrature)
- `bife` (fixed-effects logit / probit only)
- `glmer` (general mixed-effects models, including RE)
- `glmmML` (via Gauss-Hermite quadrature)
- `MCMCpack` (`MCMChlogit`)
- Various user-generated functions (e.g., [here](#)).
- Interpretation via `modelsummary` and `marginaleffects`

Stata

- `xtprobit`, `xtlogit`, `xtcloglog`
- Plus `xttrans` (transition probabilities), `quadchk` (quadrature checking), `xtrho` / `xtrhoi` (estimation of within-unit covariances)

Example: WDI “Plus”

Data from the **WDI**, plus **POLITY** and the **UCDP**:

- **ISO3** - The country's International Standards Organization (ISO) three-letter identification code.
- **Year** - The year that row of data applies to (1960=1).
- **CivilWar** - Civil conflict indicator: 1 if there was a civil conflict in that country in that year; 0 otherwise. From the **UCDP**.
- **OnsetCount** - The sum of new conflict episodes in that country / year. From **UCDP**.
- **LandArea** - Land area (sq. km).
- **PopMillions** - Population (in millions).
- **PopGrowth** - Population Growth (percent).
- **UrbanPopulation** - Urban Population (percent of total).
- **GDPPerCapita** - GDP per capita (constant 2010 \$US).
- **GDPPerCapGrowth** - GDP Per Capita Growth (percent annual).
- **PostColdWar** - 1 if Year > 1989, 0 otherwise.
- **POLITY** - The **POLITY** score of democracy/autocracy. Scaled so that 0 = most autocratic, 10 = most democratic.

$N = 215$, $\bar{T} = 64$, NT varies (due to missingness).

Model:

$$\begin{aligned} \text{Civil War}_{it} = & f[\beta_0 + \beta_1 \ln(\text{Land Area}_{it}) + \beta_2 \ln(\text{Population}_{it}) + \\ & + \beta_3 \text{Urban Population}_{it} + \beta_4 \ln(\text{GDP}_{it}) + \beta_5 \text{GDP Growth}_{it} + \\ & + \beta_6 \text{Post-Cold War}_{it} + \beta_7 \text{POLITY}_{it} + \beta_5 \text{POLITY}_{it}^2 + u_{it}] \end{aligned}$$

```
> describe(DF,skew=FALSE)
```

	vars	n	mean	sd	median	min	max	range	se
IS03*	1	13822	108.49	62.35	108.00	1.00	216.0	215.00	0.53
Year*	2	13822	32.50	18.47	32.00	1.00	64.0	63.00	0.16
YearNumeric	3	13760	1991.50	18.47	1991.50	1960.00	2023.0	63.00	0.16
country*	4	13760	108.00	62.07	108.00	1.00	215.0	214.00	0.53
CivilWar	5	9052	0.13	0.34	0.00	0.00	1.0	1.00	0.00
OnsetCount	6	9394	0.05	0.24	0.00	0.00	4.0	4.00	0.00
LandArea	7	11941	605302.93	1639812.91	107160.00	2.03	16389950.0	16389947.97	15006.31
PopMillions	8	13515	25.11	104.75	4.29	0.00	1417.2	1417.17	0.90
UrbanPopulation	9	13482	51.72	25.74	50.92	2.08	100.0	97.92	0.22
GDPPerCapita	10	10099	12088.98	19130.23	3891.83	122.88	228667.9	228545.05	190.36
GDPPerCapGrowth	11	10074	1.95	6.21	2.11	-64.43	140.5	204.91	0.06
PostColdWar	12	13760	0.53	0.50	1.00	0.00	1.0	1.00	0.00
POLITY	13	8279	5.55	3.71	6.50	0.00	10.0	10.00	0.04
POLITYSquared	14	8279	44.57	40.24	42.25	0.00	100.0	100.00	0.44

Variation

Variable	Dim	Mean	SD	Min	Max	Observations
Year	overall	1991.5	18.474	1960	2023	N = 13760
	between		0	1991.5	1991.5	n = 215
	within		18.474	1960	2023	T = 64
CivilWar	overall	0.134	0.341	0	1	N = 9052
	between		0.221	0	1	n = 172
	within		0.255	-0.783	1.117	T = 52.628
OnsetCount	overall	0.049	0.242	0	4	N = 9394
	between		0.083	0	0.597	n = 172
	within		0.227	-0.548	3.92	T = 54.616
LandArea	overall	605302.933	1639812.915	2.027	16389950	N = 11941
	between		1756124.731	2.028	16379341.333	n = 215
	within		14364.404	180581.622	688581.622	T = 55.54
PopMillions	overall	25.109	104.751	0.003	1417.173	N = 13515
	between		100.983	0.009	1104.338	n = 215
	within		28.291	-436.87	534.349	T = 62.86
UrbanPopulation	overall	51.717	25.74	2.077	100	N = 13482
	between		24.154	6.912	100	n = 214
	within		9.043	5.953	86.692	T = 63
GDPPerCapita	overall	12088.981	19130.234	122.885	228667.935	N = 10099
	between		21061.499	330.723	167809.27	n = 210
	within		6930.784	-37755.586	118168.443	T = 48.09
GDPPerCapGrowth	overall	1.952	6.215	-64.426	140.48	N = 10074
	between		1.782	-8.078	9.247	n = 212
	within		6.017	-66.481	133.991	T = 47.519
PostColdWar	overall	0.531	0.499	0	1	N = 13760
	between		0	0.531	0.531	n = 215
	within		0.499	0	1	T = 64
POLITY	overall	5.551	3.708	0	10	N = 8279
	between		2.985	0	10	n = 165
	within		2.229	-1.431	12.319	T = 50.176

Pooled Logit

```
> Logit<-glm(CivilWar~log(LandArea)+log(PopMillions)+UrbanPopulation+log(GDPPerCapita)+  
+           GDPPerCapGrowth+PostColdWar+POLITY+POLITYSquared,data=DF,family="binomial")
```

```
> summary(Logit)
```

Call:

```
glm(formula = CivilWar ~ log(LandArea) + log(PopMillions) + UrbanPopulation +  
    log(GDPPerCapita) + GDPPerCapGrowth + PostColdWar + POLITY +  
    POLITYSquared, family = "binomial", data = DF)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.81201	0.52665	-1.54	0.12312
log(LandArea)	0.00183	0.03230	0.06	0.95487
log(PopMillions)	0.66738	0.03685	18.11	< 2e-16 ***
UrbanPopulation	0.01195	0.00335	3.57	0.00036 ***
log(GDPPerCapita)	-0.52155	0.06110	-8.54	< 2e-16 ***
GDPPerCapGrowth	-0.04027	0.00651	-6.19	0.00000000061 ***
PostColdWar	-0.32313	0.08606	-3.75	0.00017 ***
POLITY	0.66857	0.06140	10.89	< 2e-16 ***
POLITYSquared	-0.06460	0.00581	-11.12	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5829.3 on 6984 degrees of freedom
Residual deviance: 4615.6 on 6976 degrees of freedom
(6837 observations deleted due to missingness)
AIC: 4634

Number of Fisher Scoring iterations: 6

Fixed Effects

```
> FELogit<-bife(CivilWar~log(LandArea)+log(PopMillions)+UrbanPopulation+log(GDPPerCapita)+
+              GDPPerCapGrowth+PostColdWar+POLITY+POLITYSquared|ISO3,data=DF,model="logit")
```

```
> summary(FELogit)
binomial - logit link
```

```
CivilWar ~ log(LandArea) + log(PopMillions) + UrbanPopulation +
  log(GDPPerCapita) + GDPPerCapGrowth + PostColdWar + POLITY +
  POLITYSquared | ISO3
```

Estimates:

	Estimate	Std. error	z value	Pr(> z)
log(LandArea)	-13.81071	8.22081	-1.68	0.093 .
log(PopMillions)	0.66428	0.29457	2.26	0.024 *
UrbanPopulation	0.01785	0.01239	1.44	0.150
log(GDPPerCapita)	-0.33280	0.17452	-1.91	0.057 .
GDPPerCapGrowth	-0.05128	0.00845	-6.07	1.3e-09 ***
PostColdWar	-0.21860	0.17884	-1.22	0.222
POLITY	0.71094	0.09360	7.60	3.1e-14 ***
POLITYSquared	-0.07364	0.00890	-8.27	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
residual deviance= 2835,
null deviance= 4408,
n= 3962, N= 83
```

```
( 6837 observation(s) deleted due to missingness )
( 3023 observation(s) deleted due to perfect classification )
```

Number of Fisher Scoring Iterations: 6

Average individual fixed effect= 172.4

Alternative Fixed Effects (using feglm)

```
> FELogit2<-feglm(CivilWar~log(LandArea)+log(PopMillions)+UrbanPopulation+log(GDPPerCapita)+
+               GDPPerCapGrowth+PostColdWar+POLITY+POLITYSquared|IS03,data=DF,family="binomial")
```

NOTES: 6,837 observations removed because of NA values (LHS: 4,770, RHS: 6,837).

77 fixed-effects (3,023 observations) removed because of only 0 (or only 1) outcomes.

```
> summary(FELogit2)
```

GLM estimation, family = binomial, Dep. Var.: CivilWar

Observations: 3,962

Fixed-effects: IS03: 83

Standard-errors: Clustered (IS03)

	Estimate	Std. Error	z value	Pr(> z)	
log(LandArea)	-13.81263	9.31735	-1.4825	0.138216796	
log(PopMillions)	0.66427	0.76041	0.8736	0.382352433	
UrbanPopulation	0.01785	0.03673	0.4860	0.626988502	
log(GDPPerCapita)	-0.33280	0.41724	-0.7976	0.425097622	
GDPPerCapGrowth	-0.05128	0.01266	-4.0508	0.000051037	***
PostColdWar	-0.21860	0.48166	-0.4538	0.649940396	
POLITY	0.71095	0.24874	2.8582	0.004260535	**
POLITYSquared	-0.07364	0.02453	-3.0018	0.002683538	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -1,417.4 Adj. Pseudo R2: 0.316066

BIC: 3,588.7 Squared Cor.: 0.400104

Random Effects

```
> RELogit<-pglm(CivilWar~log(LandArea)+log(PopMillions)+UrbanPopulation+log(GDPPerCapita)+
+               GDPPerCapGrowth+PostColdWar+POLITY+POLITYSquared|ISO3,data=DF,family=binomial,
+               effect="individual",model="random")
```

```
> summary(RELogit)
```

Maximum Likelihood estimation

Newton-Raphson maximisation, 8 iterations

Return code 3: Last step could not find a value above the current.

Boundary of parameter space?

Consider switching to a more robust optimisation method temporarily.

Log-Likelihood: -1640

10 free parameters

Estimates:

	Estimate	Std. error	t value	Pr(> t)
(Intercept)	-3.34254	2.34543	-1.43	0.15
log(LandArea)	-0.01437	0.09710	-0.15	0.88
log(PopMillions)	1.17645	0.08545	13.77	< 2e-16 ***
UrbanPopulation	0.00686	0.02895	0.24	0.81
log(GDPPerCapita)	-0.39269	0.25443	-1.54	0.12
GDPPerCapGrowth	-0.05400	0.01200	-4.50	0.0000068 ***
PostColdWar	-0.31271	0.20671	-1.51	0.13
POLITY	0.76234	0.07210	10.57	< 2e-16 ***
POLITYSquared	-0.07769	0.00663	-11.72	< 2e-16 ***
sigma	2.21027	0.12320	17.94	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Models of Civil War

	Logit	FE Logit	FES+Robust	RE Logit
Intercept	-0.81 (0.53)			-3.34 (2.35)
ln(Land Area)	0.00 (0.03)	-13.81 (8.22)	-13.81 (9.32)	-0.01 (0.10)
ln(Population)	0.67* (0.04)	0.66* (0.29)	0.66 (0.76)	1.18* (0.09)
Urban Population	0.01* (0.00)	0.02 (0.01)	0.02 (0.04)	0.01 (0.03)
ln(GDP Per Capita)	-0.52* (0.06)	-0.33 (0.17)	-0.33 (0.42)	-0.39 (0.25)
GDP Growth	-0.04* (0.01)	-0.05* (0.01)	-0.05* (0.01)	-0.05* (0.01)
Post-Cold War	-0.32* (0.09)	-0.22 (0.18)	-0.22 (0.48)	-0.31 (0.21)
POLITY	0.67* (0.06)	0.71* (0.09)	0.71* (0.25)	0.76* (0.07)
POLITY Squared	-0.06* (0.01)	-0.07* (0.01)	-0.07* (0.02)	-0.08* (0.01)
Estimated Sigma				2.21* (0.12)
AIC	4633.63			3299.59
BIC	4695.30			
Log Likelihood	-2307.82	-1417.42	-1417.42	-1639.80
Deviance	4615.63	2834.83	2834.83	
Num. obs.	6985	3962	3962	
Num. groups: ISO3			83	
Pseudo R ²			0.32	

* $p < 0.05$

Models For Event Counts

Properties:

- Discrete / integer-values
- Non-negative
- “Cumulative”

Motivation:

$$\text{Arrival Rate} = \lambda$$

$$\Pr(\text{Event})_{t,t+h} = \lambda h$$

$$\Pr(\text{No Event})_{t,t+h} = 1 - \lambda h$$

$$\begin{aligned}\Pr(Y_t = y) &= \frac{\exp(-\lambda h) \lambda h^y}{y!} \\ &= \frac{\exp(-\lambda) \lambda^y}{y!}\end{aligned}$$

Poisson: Assumptions and Motivations

Three key assumptions:

- No Simultaneous Events
- Constant Arrival Rate
- Independent Event Arrivals

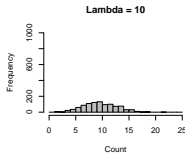
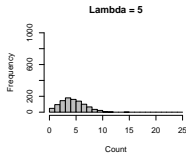
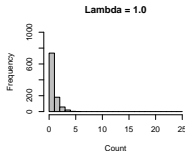
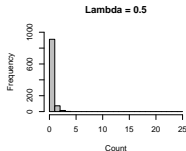
Another motivation: For M independent Bernoulli trials with (sufficiently small) probability of success π and where $M\pi \equiv \lambda > 0$,

$$\begin{aligned}\Pr(Y_i = y) &= \lim_{M \rightarrow \infty} \left[\binom{M}{y} \left(\frac{\lambda}{M}\right)^y \left(1 - \frac{\lambda}{M}\right)^{M-y} \right] \\ &= \frac{\lambda^y \exp(-\lambda)}{y!}\end{aligned}$$

Poisson: Characteristics

Characteristics:

- Discrete
- $E(Y) = \text{Var}(Y) = \lambda$
- Is not preserved under affine transformations...
- For $X \sim \text{Poisson}(\lambda_X)$ and $Y \sim \text{Poisson}(\lambda_Y)$, $Z = X + Y \sim \text{Poisson}(\lambda_X + \lambda_Y)$
iff X and Y are independent but
- ...same is not true for differences.
- $\lambda \rightarrow \infty \iff Y \sim N$



Suppose

$$E(Y_i) \equiv \lambda_i = \exp(\mathbf{X}_i\beta)$$

then

$$\Pr(Y_i = y | \mathbf{X}_i, \beta) = \frac{\exp[-\exp(\mathbf{X}_i\beta)][\exp(\mathbf{X}_i\beta)]^y}{y!}$$

with likelihood:

$$L = \prod_{i=1}^N \frac{\exp[-\exp(\mathbf{X}_i\beta)][\exp(\mathbf{X}_i\beta)]^{Y_i}}{Y_i!}$$

and log-likelihood:

$$\ln L = \sum_{i=1}^N [-\exp(\mathbf{X}_i\beta) + Y_i\mathbf{X}_i\beta - \ln(Y_i!)]$$

Event Counts: Unit Effects

The Poisson model:

$$Y_{it} \sim \text{Poisson}(\mu_{it} = \alpha_i \lambda_{it})$$

with $\lambda_{it} = \exp(\mathbf{X}_{it}\beta)$ implies:

$$\begin{aligned} E(Y_{it} \mid \mathbf{X}_{it}, \alpha_i) &= \mu_{it} \\ &= \alpha_i \exp(\mathbf{X}_{it}\beta) \\ &= \exp(\delta_i + \mathbf{X}_{it}\beta) \end{aligned}$$

where $\delta_i = \ln(\alpha_i)$.

Fixed-Effects Poisson:

- ...has no “incidental parameters” problem (see e.g. Cameron and Trivedi, pp. 281-2)
- This means “brute force” approach works
- Fitted via `glmmML` in R, `xtpoisson` (and `xtnbreg`) in Stata

The Poisson with random effects is:

$$\begin{aligned}\Pr(Y_{i1} = y_{i1}, \dots, Y_{iT} = y_{iT}) &= \int_0^\infty \Pr(Y_{i1} = y_{i1}, \dots, Y_{iT} = y_{iT}) f(\alpha_i) d\alpha_i \\ &= \int_0^\infty \left[\prod_{t=1}^T \Pr(Y_{it} | \alpha_i) \right] f(\alpha_i) d\alpha_i\end{aligned}$$

- Simplest to assume $\alpha_i \sim \Gamma(\theta)$
- Yields a model with $E(Y_{it}) = \lambda_{it}$ and $\text{Var}(Y_{it}) = \lambda_{it} + \frac{\lambda_{it}^2}{\theta}$
- Via `glmmML` or `glmer` in R, or `xtpois`, `re` in Stata
- \exists random effects negative binomial too...

R:

- Tobit = `censReg` (in **`censReg`**)
- Poisson (random effects) = `glmmML` in **`glmmML`** or `glmer` in **`lme4`**
- Poisson (fixed effects) = `glmmML` or “brute force”
- Poisson + negative binomial (FE, RE) = `pglm`

Stata:

- Tobit = `xttobit` (re only)
- Poisson / negative binomial = `xtpoisson`, `xtnbreg` (both with `fe`, `re` options)

Conflict Onsets: Pooled Poisson

```
> xtabs(~DF$OnsetCount)
DF$OnsetCount
  0    1    2    3    4
8981 375   30    7    1

> Poisson<-glm(OnsetCount~log(LandArea)+log(PopMillions)+UrbanPopulation+log(GDPPerCapita)+
+             GDPPerCapGrowth+PostColdWar+POLITY+POLITYSquared,data=DF,family="poisson")

> summary(Poisson)

Call:
glm(formula = OnsetCount ~ log(LandArea) + log(PopMillions) +
    UrbanPopulation + log(GDPPerCapita) + GDPPerCapGrowth + PostColdWar +
    POLITY + POLITYSquared, family = "poisson", data = DF)

Coefficients:
              Estimate Std. Error z value    Pr(>|z|)
(Intercept)   -1.99988    0.72467   -2.76    0.00579 **
log(LandArea)    0.06397    0.04707    1.36    0.17417
log(PopMillions) 0.42562    0.04573    9.31 < 2e-16 ***
UrbanPopulation  0.00804    0.00472    1.70    0.08879 .
log(GDPPerCapita) -0.47969    0.08043   -5.96 0.0000000025 ***
GDPPerCapGrowth -0.03581    0.00668   -5.36 0.0000000817 ***
PostColdWar      0.25159    0.12064    2.09    0.03703 *
POLITY           0.30534    0.08363    3.65    0.00026 ***
POLITYSquared    -0.03366    0.00799   -4.21 0.0000254512 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2371.5  on 6984  degrees of freedom
Residual deviance: 1931.5  on 6976  degrees of freedom
(6837 observations deleted due to missingness)
AIC: 2679

Number of Fisher Scoring iterations: 6
```

Fixed Effects Poisson

```
> FEpoisson<-pglm(OnsetCount~log(LandArea)+log(PopMillions)+UrbanPopulation+
+               log(GDPPerCapita)+GDPPerCapGrowth+PostColdWar+POLITY+
+               POLITYSquared,data=DF,family="poisson",effect="individual",
+               model="within")
```

```
> summary(FEpoisson)
```

Maximum Likelihood estimation

Newton-Raphson maximisation, 3 iterations

Return code 8: successive function values within relative tolerance limit (reltol)

Log-Likelihood: -1013

8 free parameters

Estimates:

	Estimate	Std. error	t value	Pr(> t)
log(LandArea)	-2.82014	2.86711	-0.98	0.32530
log(PopMillions)	0.61597	0.31877	1.93	0.05332 .
UrbanPopulation	-0.04529	0.01338	-3.38	0.00071 ***
log(GDPPerCapita)	-0.10765	0.14544	-0.74	0.45919
GDPPerCapGrowth	-0.02872	0.00686	-4.19	0.0000282 ***
PostColdWar	0.47277	0.19572	2.42	0.01571 *
POLITY	0.51291	0.10812	4.74	0.0000021 ***
POLITYSquared	-0.05185	0.01061	-4.89	0.0000010 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Equivalent Fixed Effects Poisson (using feglm)

```
> FEpoisson2<-feglm(OnsetCount~log(LandArea)+log(PopMillions)+UrbanPopulation+
+                   log(GDPPerCapita)+GDPPerCapGrowth+PostColdWar+POLITY+
+                   POLITYSquared|IS03,data=DF,family="poisson")
```

NOTES: 6,837 observations removed because of NA values (LHS: 4,428, RHS: 6,837).
67 fixed-effects (2,495 observations) removed because of only 0 outcomes.

```
> summary(FEpoisson2,cluster="IS03")
```

GLM estimation, family = poisson, Dep. Var.: OnsetCount

Observations: 4,490

Fixed-effects: IS03: 93

Standard-errors: Clustered (IS03)

	Estimate	Std. Error	z value	Pr(> z)
log(LandArea)	-2.82014	3.675097	-0.7674	0.4428652862
log(PopMillions)	0.61597	0.347129	1.7745	0.0759873451 .
UrbanPopulation	-0.04529	0.019638	-2.3061	0.0211045803 *
log(GDPPerCapita)	-0.10765	0.153162	-0.7029	0.4821364556
GDPPerCapGrowth	-0.02872	0.006582	-4.3638	0.0000127806 ***
PostColdWar	0.47277	0.295534	1.5997	0.1096591753
POLITY	0.51291	0.111671	4.5931	0.0000043681 ***
POLITYSquared	-0.05185	0.011586	-4.4757	0.0000076159 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -1,148.7 Adj. Pseudo R2: 0.093386

BIC: 3,146.7 Squared Cor.: 0.163105

Random Effects Poisson

```
> REPoisson<-pglm(OnsetCount~log(LandArea)+log(PopMillions)+UrbanPopulation+
+               log(GDPPerCapita)+GDPPerCapGrowth+PostColdWar+POLITY+
+               POLITYSquared,data=DF,family="poisson",effect="individual",
+               model="random")
```

```
> summary(REPoisson)
```

Maximum Likelihood estimation

Newton-Raphson maximisation, 4 iterations

Return code 8: successive function values within relative tolerance limit (reltol)

Log-Likelihood: -1283

10 free parameters

Estimates:

	Estimate	Std. error	t value	Pr(> t)	
(Intercept)	-3.47867	1.04013	-3.34	0.00082	***
log(LandArea)	0.05347	0.07261	0.74	0.46146	
log(PopMillions)	0.44330	0.07919	5.60	0.000000022	***
UrbanPopulation	-0.00471	0.00639	-0.74	0.46120	
log(GDPPerCapita)	-0.22335	0.10250	-2.18	0.02934	*
GDPPerCapGrowth	-0.03367	0.00683	-4.93	0.000000828	***
PostColdWar	0.28226	0.12865	2.19	0.02823	*
POLITY	0.45969	0.09613	4.78	0.000001734	***
POLITYSquared	-0.05071	0.00931	-5.45	0.000000050	***
sigma	1.77551	0.44025	4.03	0.000055069	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Negative Binomial: Fixed Effects

```
> FENegBin2<-fenegbin(OnsetCount~log(LandArea)+log(PopMillions)+  
+      UrbanPopulation+log(GDPPerCapita)+  
+      GDPPerCapGrowth+PostColdWar+POLITY+  
+      POLITYSquared|IS03,data=DF)
```

NOTES: 6,837 observations removed because of NA values (LHS: 4,428, RHS: 6,837).
67 fixed-effects (2,495 observations) removed because of only 0 outcomes.

Very high value of theta (10000). There is no sign of overdispersion, you may consider a Poisson model.

Warning message:

[femlm]: The information matrix is singular: presence of collinearity.

Models of Civil War Onset Counts

	Poisson	FE Poisson	RE Poisson
Intercept	-2.00* (0.72)		-3.48* (1.04)
ln(Land Area)	0.06 (0.05)	-2.82 (2.87)	0.05 (0.07)
ln(Population)	0.43* (0.05)	0.62 (0.32)	0.44* (0.08)
Urban Population	0.01 (0.00)	-0.05* (0.01)	-0.00 (0.01)
ln(GDP Per Capita)	-0.48* (0.08)	-0.11 (0.15)	-0.22* (0.10)
GDP Growth	-0.04* (0.01)	-0.03* (0.01)	-0.03* (0.01)
Post-Cold War	0.25* (0.12)	0.47* (0.20)	0.28* (0.13)
POLITY	0.31* (0.08)	0.51* (0.11)	0.46* (0.10)
POLITY Squared	-0.03* (0.01)	-0.05* (0.01)	-0.05* (0.01)
Estimated Sigma			1.78* (0.44)
AIC	2679.09	2041.78	2585.34
BIC	2740.75		
Log Likelihood	-1330.54	-1012.89	-1282.67
Deviance	1931.52		
Num. obs.	6985		

* $p < 0.05$

Wrap-Up: Some Useful Packages

- `pglm`
 - Workhorse package for panel (FE, RE, BE) GLMs
 - Binary + ordered logit/probit, Poisson / negative binomial
 - Discussed + used extensively in Croissant and Millo (2018) *Panel Data Econometrics with R*
 - The one thing it won't (apparently) do is fixed-effects, binary-response models...
- `fixest`
 - Fast / efficient fitting of FE models
 - Fits linear models, logit, Poisson, and negative binomial
 - Includes easy coefficient plots & tables; simple multi-threading; built-in "robust" S.E.s
- `alpaca`
 - Fast / efficient fitting of GLMs with high-dimensional fixed effects
 - *Includes bias correction for incidental parameters after binary-response models*
 - Also includes useful panel data simulation routines + average partial effects

Generalized Estimating Equations

Linear-normal model is:

$$Y_i = \mu_i + u_i$$

with:

$$\mu_i = \mathbf{X}_i \boldsymbol{\beta}.$$

Generalize:

$$g(\mu_i) = \mathbf{X}_i \boldsymbol{\beta}$$

and:

$$Y_i \sim \text{i.i.d. } F[\mu_i, \mathbf{V}_i].$$

“Score” equations:

$$\mathbf{U}(\beta) = \sum_{i=1}^N \mathbf{D}_i' \mathbf{V}_i^{-1} [Y_i - \mu_i] = \mathbf{0}.$$

with:

- $\mathbf{D}_i = \frac{\partial \mu_i}{\partial \beta}$,
- $\mathbf{V}_i = \frac{h(\mu_i)}{\phi}$, and
- $(Y_i - \mu_i) \approx$ a “residual.”
- Known as “quasi-likelihood” (e.g. Wedderburn 1974 *Biometrika*).

Now suppose:

$$Y_{it} = \mu_{it} + u_{it}$$

where

- $i \in \{1, \dots, N\}$ are i.i.d. “units,”
- $t \in \{1, \dots, T\}$, $T > 1$ are “time points,”
- we want $g(\mu_{it}) = \mathbf{X}_{it}\beta$.

Key issue: Accounting for (conditional) dependence in Y over time.

Full joint distributions over T are hard. But...

Define:

$$\mathbf{R}_i(\boldsymbol{\alpha})_{T \times T} = \begin{pmatrix} 1.0 & \alpha_{12} & \cdots & \alpha_{1,T} \\ \alpha_{21} & 1.0 & \cdots & \alpha_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{T,1} & \cdots & \alpha_{T,T-1} & 1.0 \end{pmatrix},$$

→ “working correlation” matrix.

- Completely defined by $\boldsymbol{\alpha}$,
- Structure specified by the analyst.

Liang and Zeger (1986): We can decompose the variance of Y_{it} as:

$$\mathbf{V}_i = \text{diag}(\mathbf{V}_i^{\frac{1}{2}}) \mathbf{R}_i(\boldsymbol{\alpha}) \text{diag}(\mathbf{V}_i^{\frac{1}{2}})$$

With a standard GLM assumption about the mean and variance, this is:

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

where

$$\mathbf{A}_i = \begin{pmatrix} h(\mu_{i1}) & 0 & \cdots & 0 \\ 0 & h(\mu_{i2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & h(\mu_{iT}) \end{pmatrix}$$

$\mathbf{V}_i = \text{Var}(Y_{it} | \mathbf{X}_{it}, \beta)$ has two parts:

- $\mathbf{A}_i = \underline{\text{unit-level}}$ variation,
- $\mathbf{R}_i(\alpha) = \text{within-unit } \underline{\text{temporal}}$ variation.

Specifying $\mathbf{R}_i(\alpha)$

Independent:

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1.0 & 0 & \cdots & 0 \\ 0 & 1.0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1.0 \end{pmatrix}$$

- Assumes no within-unit temporal correlation.
- Equivalent to GLM on pooled data.

Exchangeable:

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1.0 & \alpha & \cdots & \alpha \\ \alpha & 1.0 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \cdots & \alpha & 1.0 \end{pmatrix}$$

- One free parameter in $\mathbf{R}_i(\alpha)$ ($\alpha_{ts} = \alpha \forall t \neq s$)
- Temporal correlation within units is constant across time points.
- Akin (in some respects) to a random-effects model...

Specifying $\mathbf{R}_i(\alpha)$

$AR(p)$ (e.g., $AR(1)$):
$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1.0 & \alpha & \alpha^2 & \cdots & \alpha^{T-1} \\ \alpha & 1.0 & \alpha & \cdots & \alpha^{T-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha^{T-1} & \cdots & \alpha^2 & \alpha & 1.0 \end{pmatrix}$$

- One free parameter in $\mathbf{R}_i(\alpha)$ ($\alpha_{ts} = \alpha^{|t-s|} \forall t \neq s$).
- Conditional within-unit correlation an exponential function of the lag.

$Stationary(p)$:
$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1.0 & \alpha_1 & \cdots & \alpha_p & 0 & \cdots & 0 \\ \alpha_1 & 1.0 & \alpha_1 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \alpha_p & \cdots & \alpha_1 & 1.0 \end{pmatrix}$$

- AKA “banded,” or “ p -dependent.”
- $p \leq T - 1$ free parameters in $\mathbf{R}_i(\alpha)$.
- Conditional within-unit correlation an exponential function of the lag, up to lag p , and zero thereafter.

Specifying $\mathbf{R}_i(\alpha)$

Unstructured: $\mathbf{R}_i(\alpha) = \begin{pmatrix} 1.0 & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1,T-1} \\ \alpha_{12} & 1.0 & \alpha_{23} & \cdots & \alpha_{2,T-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha_{1,T-1} & \alpha_{2,T-1} & \cdots & \alpha_{T-1,T-1} & 1.0 \end{pmatrix}$

- $\frac{T(T-1)}{2}$ free parameters in $\mathbf{R}_i(\alpha)$.
- Conditional within-unit correlation is completely data-dependent.

Score equations:

$$\mathbf{U}_{GEE}(\boldsymbol{\beta}_{GEE}) = \sum_{i=1}^N \mathbf{D}_i' \left[\frac{(\mathbf{A}_i^{\frac{1}{2}}) \mathbf{R}_i(\boldsymbol{\alpha}) (\mathbf{A}_i^{\frac{1}{2}})}{\phi} \right]^{-1} [Y_i - \mu_i] = \mathbf{0}$$

Two-step estimation:

- For fixed values of $\boldsymbol{\alpha}_s$ and ϕ_s at iteration s , use Newton scoring to estimate $\hat{\boldsymbol{\beta}}_s$,
- Use $\hat{\boldsymbol{\beta}}_s$ to calculate standardized residuals $(Y_i - \hat{\mu}_i)_s$, from which consistent estimates of $\boldsymbol{\alpha}_{s+1}$ and ϕ_{s+1} can be estimated.

Liang & Zeger (1986):

$$\hat{\beta}_{GEE} \underset{N \rightarrow \infty}{\sim} \mathbf{N}(\beta, \Sigma).$$

For $\hat{\Sigma}$, two options:

$$\hat{\Sigma}_{\text{Model}} = N \left(\sum_{i=1}^N \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right)$$

$$\hat{\Sigma}_{\text{Robust}} = N \left(\sum_{i=1}^N \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right)^{-1} \left(\sum_{i=1}^N \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{S}}_i \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right) \left(\sum_{i=1}^N \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right)^{-1}$$

where $\hat{\mathbf{S}}_i = (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$.

Inference (aka, magic!)

- $\hat{\Sigma}_{\text{Model}}$
 - Requires that $\mathbf{R}_i(\alpha)$ be “correct” for consistency.
 - Is slightly more efficient than $\hat{\Sigma}_{\text{Robust}}$ if so.
- $\hat{\Sigma}_{\text{Robust}}$
 - Is consistent *even if* $\mathbf{R}_i(\alpha)$ is misspecified.
 - Is slightly less efficient than $\hat{\Sigma}_{\text{Model}}$ if $\mathbf{R}_i(\alpha)$ is correct.

Moral: Use $\hat{\Sigma}_{\text{Robust}}$.

GEEs:

- Are a straightforward variation on GLMs, and so
- Can be applied to a range of data types (continuous, binary, count, proportions, etc.),
- Yield robustly consistent point estimates of β s,
- Account for within-unit correlation in an informed way, but also
- Yield consistent inferences even if the correlation is misspecified.

Practical Issues: Model Interpretation

- In general, GEEs = GLMs.
- GEEs are *marginal* models, so:
 - $\hat{\beta}$ s have an interpretation as average / total effects.
 - Estimates / effect sizes generally be smaller than conditional (e.g. fixed/random) effects models.
 - E.g., for logit, $\hat{\beta}_M \approx \frac{\hat{\beta}_C}{\sqrt{1+0.35\sigma_\eta^2}}$, where $\sigma_\eta^2 > 0$ is the variance of the unit effects.
 - See (e.g.) [Gardiner et al. \(2009\)](#) or [Koper and Manseau \(2009\)](#) for expositions.

Practical Issues: Specifying $\mathbf{R}_i(\alpha)$

- Has been called “more art than science.”
- Pointers:
 - Choose based on *substance* of the problem.
 - Remember that $\mathbf{R}_i(\alpha)$ is conditional on \mathbf{X} , $\hat{\beta}$.
 - Consider unstructured when T is small and N large.
 - Try different ones, and compare.
- In general, it shouldn't matter terribly much...

Software	Command(s)/Package(s)
R	gee / geepack / geeM / multgeeB / orth / repolr
Stata	xtgee / xtlogit / xtprobit / xtpois / etc.
SAS	genmod (w/ repeated)

- Generally follow GLMs (specify “family” + “link”)
- Certain combinations not possible/recommended
- Estimation: Fisher scoring, MLE, etc. (MCMC?)

From the geepack manual:

Warning

Use "unstructured" correlation structure only with great care. (It may cause R to crash).

Civil War Redux... GEE: Independence

```
> GEE.ind<-geeglm(CivilWar~log(LandArea)+log(PopMillions)+UrbanPopulation+
+               log(GDPPerCapita)+GDPPerCapGrowth+PostColdWar+POLITY+
+               POLITYSquared,data=DF,id=IS03,family="binomial",
+               corstr="independence")
```

```
> summary(GEE.ind)
```

Call:

```
geeglm(formula = CivilWar ~ log(LandArea) + log(PopMillions) +
        UrbanPopulation + log(GDPPerCapita) + GDPPerCapGrowth + PostColdWar +
        POLITY + POLITYSquared, family = "binomial", data = DF, id = IS03,
        corstr = "independence")
```

Coefficients:

	Estimate	Std.err	Wald	Pr(> W)	
(Intercept)	-0.81201	2.02183	0.16	0.68796	
log(LandArea)	0.00183	0.12331	0.00	0.98817	
log(PopMillions)	0.66738	0.15618	18.26	0.000019	***
UrbanPopulation	0.01195	0.01405	0.72	0.39489	
log(GDPPerCapita)	-0.52155	0.25237	4.27	0.03877	*
GDPPerCapGrowth	-0.04027	0.01302	9.57	0.00198	**
PostColdWar	-0.32313	0.26138	1.53	0.21637	
POLITY	0.66857	0.21145	10.00	0.00157	**
POLITYSquared	-0.06460	0.01951	10.97	0.00093	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation structure = independence

Estimated Scale Parameters:

	Estimate	Std.err
(Intercept)	0.805	0.295

Number of clusters: 160 Maximum cluster size: 57

GEE: Exchangeable

```
> GEE.exc<-geeglm(CivilWar~log(LandArea)+log(PopMillions)+UrbanPopulation+log(GDPPerCapita)+GDPPerCapGrowth+  
+ PostColdWar+POLITY+POLITYSquared,data=DF,id=IS03,family="binomial",corstr="exchangeable")
```

```
> summary(GEE.exc)
```

Call:

```
geeglm(formula = CivilWar ~ log(LandArea) + log(PopMillions) +  
UrbanPopulation + log(GDPPerCapita) + GDPPerCapGrowth + PostColdWar +  
POLITY + POLITYSquared, family = "binomial", data = DF, id = IS03,  
corstr = "exchangeable")
```

Coefficients:

	Estimate	Std.err	Wald	Pr(> W)
(Intercept)	-2.67543	2.04945	1.70	0.19174
log(LandArea)	0.03410	0.15498	0.05	0.82585
log(PopMillions)	0.55616	0.16182	11.81	0.00059 ***
UrbanPopulation	0.00542	0.01168	0.22	0.64247
log(GDPPerCapita)	-0.22187	0.17520	1.60	0.20538
GDPPerCapGrowth	-0.03599	0.00911	15.62	0.000078 ***
PostColdWar	-0.14495	0.23381	0.38	0.53528
POLITY	0.55143	0.17124	10.37	0.00128 **
POLITYSquared	-0.05620	0.01675	11.26	0.00079 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation structure = exchangeable

Estimated Scale Parameters:

	Estimate	Std.err
(Intercept)	0.729	0.178

Link = identity

Estimated Correlation Parameters:

	Estimate	Std.err
alpha	0.342	0.112

Number of clusters: 160 Maximum cluster size: 57

GEE: AR(1)

```
> GEE.ar1<-geeglm(CivilWar~log(LandArea)+log(PopMillions)+UrbanPopulation+log(GDPPerCapita)+GDPPerCapGrowth+
+               PostColdWar+POLITY+POLITYSquared,data=DF,id=ISO3,family="binomial",corstr="ar1")

> summary(GEE.ar1)

Call:
geeglm(formula = CivilWar ~ log(LandArea) + log(PopMillions) +
        UrbanPopulation + log(GDPPerCapita) + GDPPerCapGrowth + PostColdWar +
        POLITY + POLITYSquared, family = "binomial", data = DF, id = ISO3,
        corstr = "ar1")

Coefficients:
              Estimate Std.err Wald Pr(>|W|)
(Intercept)   -1.14508   3.06369  0.14   0.709
log(LandArea)    0.08037   0.21013  0.15   0.702
log(PopMillions) 0.37575   0.21730  2.99   0.084 .
UrbanPopulation -0.00320   0.01795  0.03   0.858
log(GDPPerCapita) -0.35783   0.27098  1.74   0.187
GDPPerCapGrowth -0.01643   0.00793  4.30   0.038 *
PostColdWar     0.20467   0.24582  0.69   0.405
POLITY          0.19608   0.12482  2.47   0.116
POLITYSquared   -0.02126   0.01309  2.64   0.104
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation structure = ar1
Estimated Scale Parameters:

              Estimate Std.err
(Intercept)    0.818   0.374
Link = identity

Estimated Correlation Parameters:
              Estimate Std.err
alpha        0.922   0.039
Number of clusters: 160 Maximum cluster size: 57
```

GEE: Unstructured (2013-2017)

```
> GEE.unstr<-geeglm(CivilWar~log(LandArea)+log(PopMillions)+UrbanPopulation+log(GDPPerCapita)+GDPPerCapGrowth+POLITY+
+ POLITYSquared,data=DF5,id=IS03,family="binomial",corstr="unstructured")
```

```
> summary(GEE.unstr)
```

Call:

```
geeglm(formula = CivilWar ~ log(LandArea) + log(PopMillions) +
  UrbanPopulation + log(GDPPerCapita) + GDPPerCapGrowth + POLITY +
  POLITYSquared, family = "binomial", data = DF5, id = IS03,
  corstr = "unstructured")
```

Coefficients:

	Estimate	Std.err	Wald	Pr(> W)
(Intercept)	-1.9922	3.1555	0.40	0.52782
log(LandArea)	0.1442	0.1930	0.56	0.45489
log(PopMillions)	0.8840	0.2488	12.62	0.00038 ***
UrbanPopulation	0.0354	0.0171	4.27	0.03884 *
log(GDPPerCapita)	-0.8469	0.3089	7.52	0.00611 **
GDPPerCapGrowth	-0.0125	0.0297	0.18	0.67372
POLITY	0.5091	0.4071	1.56	0.21111
POLITYSquared	-0.0588	0.0359	2.69	0.10094

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation structure = unstructured

Estimated Scale Parameters:

	Estimate	Std.err
(Intercept)	0.675	0.813

Link = identity

Estimated Correlation Parameters:

	Estimate	Std.err
alpha.1:2	0.389	0.490
alpha.1:3	0.409	0.515
alpha.1:4	0.341	0.433
alpha.1:5	0.334	0.423
alpha.2:3	0.728	0.849
alpha.2:4	0.276	0.354
alpha.2:5	0.503	0.588
alpha.3:4	0.396	0.508
alpha.3:5	0.741	0.875
alpha.4:5	0.432	0.548

Number of clusters: 159 Maximum cluster size: 5

GEE Model Comparison

GEE Models of Civil War Onset

	Independence	Exchangeable	AR(1)	Unstructured (2013-17)
(Intercept)	-0.812 (2.022)	-2.675 (2.049)	-1.145 (3.064)	-1.992 (3.155)
ln(Land Area)	0.002 (0.123)	0.034 (0.155)	0.080 (0.210)	0.144 (0.193)
ln(Population)	0.667*** (0.156)	0.556*** (0.162)	0.376+ (0.217)	0.884*** (0.249)
Urban Population	0.012 (0.014)	0.005 (0.012)	-0.003 (0.018)	0.035* (0.017)
ln(GDP Per Capita)	-0.522* (0.252)	-0.222 (0.175)	-0.358 (0.271)	-0.847** (0.309)
GDP Growth	-0.040** (0.013)	-0.036*** (0.009)	-0.016* (0.008)	-0.013 (0.030)
Post-Cold War	-0.323 (0.261)	-0.145 (0.234)	0.205 (0.246)	
POLITY	0.669** (0.211)	0.551** (0.171)	0.196 (0.125)	0.509 (0.407)
POLITY Squared	-0.065*** (0.020)	-0.056*** (0.017)	-0.021 (0.013)	-0.059 (0.036)
<i>NT</i>	6985	6985	6985	790

Note: + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

GEEs are:

- Robust
- Flexible
- Extensible beyond panel/TSCS context

Appendix: Discrete-Time Survival Models

Survival models:

- ...are models for *time-to-event data*.
- ...have their roots in biostats/epidemiology, plus engineering, sociology, economics.
- Examples...
 - Political careers, confirmation durations, position-taking, bill cosponsorship, campaign contributions, policy innovation/adoption, etc.
 - Cabinet/government durations, length of civil wars, coalition durability, etc.
 - War duration, peace duration, alliance longevity, length of trade agreements, etc.
 - Strike durations, work careers (including promotions, firings, etc.), criminal careers, marriage and child-bearing behavior, etc.

Time-To-Event Data

Characteristics:

- Discrete events (i.e., not continuous),
- Take place over time,
- May not (or never) experience the event (i.e., possibility of censoring).

Terminology:

- Y_i = the duration until the event occurs,
- Z_i = the duration until the observation is “censored”
- T_i = $\min\{Y_i, Z_i\}$,
- C_i = 0 if observation i is *censored*, 1 if it is not.

Density:

$$f(t) = \Pr(T_i = t)$$

CDF:

$$\Pr(T_i \leq t) \equiv F(t) = \int_0^t f(t) dt$$

Survival function:

$$\begin{aligned}\Pr(T_i \geq t) \equiv S(t) &= 1 - F(t) \\ &= 1 - \int_0^t f(t) dt\end{aligned}$$

Hazard:

$$\begin{aligned}\Pr(T_i = t | T_i \geq t) \equiv h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{f(t)}{1 - \int_0^t f(t) dt}\end{aligned}$$

Grouped-Data Survival Approaches

Model:

$$\Pr(C_{it} = 1) = f(\mathbf{X}_{it}\beta)$$

Advantages:

- Easily estimated, interpreted and understood
- Natural interpretations:
 - $\hat{\beta}_0 \approx$ “baseline hazard”
 - Covariates shift this up or down.
- Can incorporate data on time-varying covariates
- Lots of software

Potential Disadvantages:

- Requires time-varying data
- Must deal with time dependence explicitly

Temporal Issues in Grouped-Data Models

(Implicit) “Baseline” hazard:

$$h_0(t) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

→ No temporal dependence / “flat” hazard

Time trend:

$$\Pr(Y_{it} = 1) = f(\mathbf{X}_{it}\beta + \gamma T_{it})$$

- $\hat{\gamma} > 0 \rightarrow$ rising hazard
- $\hat{\gamma} < 0 \rightarrow$ declining hazard
- $\hat{\gamma} = 0 \rightarrow$ “flat” (exponential) hazard

Variants/extensions: Polynomials...

$$\Pr(Y_{it} = 1) = f(\mathbf{X}_{it}\beta + \gamma_1 T_{it} + \gamma_2 T_{it}^2 + \gamma_3 T_{it}^3 + \dots)$$

Temporal Issues in Grouped-Data Models

“Time dummies”:

$$\Pr(Y_{it} = 1) = f[\mathbf{X}_{it}\beta + \alpha_1 I(T_{i1}) + \alpha_2 I(T_{i2}) + \dots + \alpha_{t_{\max}} I(T_{it_{\max}})]$$

→ Beck, Katz, and Tucker's (1998) cubic splines; might also use:

- Fractional polynomials
- Smoothed duration
- Loess/lowess fits
- Other splines (B-splines, P-splines, natural splines, etc.)