GSERM - St. Gallen 2025

Analyzing Panel Data

June 20, 2025

Start with:

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

$$\begin{aligned} Y_i &= 0 \text{ if } Y_i^* < 0 \\ Y_i &= 1 \text{ if } Y_i^* \geq 0 \end{aligned}$$

So:

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

"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"

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

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

Likelihoods

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

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

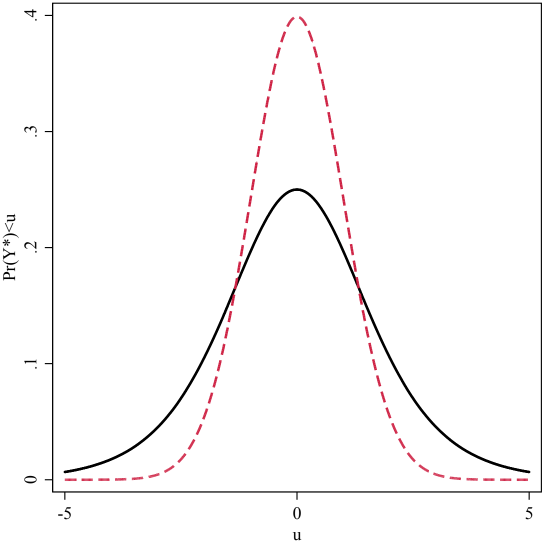
$$\begin{aligned}\ln L \quad = \quad & \sum_{i=1}^{N} Y_i \ln\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right) + \\ & (1-Y_i)\ln\left[1-\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right)\right]\end{aligned}$$

Probit...

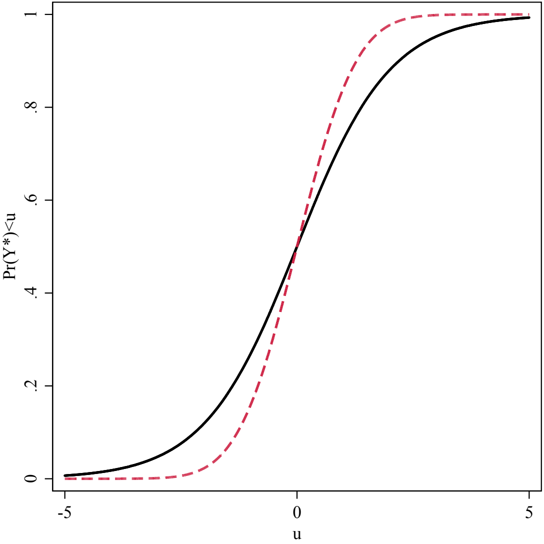
$$\Pr(u) \equiv \phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

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

Standard Normal and Logistic PDFs



Standard Normal and Logistic CDFs



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

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

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

One-way unit effects:

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

So, think about logit first:

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

<u>Incidental Parameters</u>:

- Nonlinearity $\rightarrow$ inconsistency in both $\hat{\alpha}$s <u>and</u> $\hat{\boldsymbol{\beta}}$.
- 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:

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

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

<u>The Point</u>:

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

Things to bear in mind:

- Fixed effects = no estimates for $\boldsymbol{\beta}_b$
- Interpretation: per logit, but $\mid \hat{\alpha}_i$.
- 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:

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

with:

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

with $\eta_{it} \sim$ 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 $\sigma_\alpha^2 = \left(\frac{\rho}{1-\rho}\right)$.

Random Effects Variants

Probit:

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

Logit:

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

Solution?

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

Points:

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

Software

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` - Popluation (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 and Data

Model:

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

```
> describe(DF,skew=FALSE)
                vars     n      mean         sd    median     min        max       range       se
ISO3*              1 14037    108.48      62.35    108.00    1.00      216.0      215.00     0.53
Year*              2 14037     32.99      18.76     33.00    1.00       65.0       64.00     0.16
YearNumeric        3 13975   1992.00      18.76   1992.00 1960.00     2024.0       64.00     0.16
country*           4 13975    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 12068 596744.44 1643440.14 100915.00    2.03 16389950.0 16389947.97 14960.16
PopMillions        8 13730     25.35     105.43      4.31    0.00     1438.1     1438.07     0.90
UrbanPopulation    9 13696     51.89      25.74     51.17    2.08      100.0       97.92     0.22
GDPPerCapita      10 11040  12313.79   19265.74   3933.86  122.68   224582.5   224459.77   183.36
GDPPerCapGrowth   11 10970      1.89       6.49      2.09  -64.42      140.5      204.91     0.06
PostColdWar       12 13975      0.54       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	1992	18.762	1960	2024	$NT = 13975$
	between		0	1992	1992	$N = 215$
	within		18.762	1960	2024	$T = 65$
Civil War	overall	0.134	0.341	0	1	$NT = 9052$
	between		0.221	0	1	$N = 172$
	within		0.255	-0.783	1.117	$T = 52.628$
Onset Count	overall	0.049	0.242	0	4	$NT = 9394$
	between		0.083	0	0.597	$N = 172$
	within		0.227	-0.548	3.92	$T = 54.616$
Land Area	overall	596744.441	1643440.135	2.027	16389950	$NT = 12068$
	between		1754374.985	2.029	16379261.613	$N = 215$
	within		5283.701	425560.57	652779.912	$T = 56.13$
Pop. Millions	overall	25.346	105.43	0.003	1438.07	$NT = 13730$
	between		101.488	0.008	1109.125	$N = 215$
	within		29.008	-450.719	551.361	$T = 63.86$
Urban Population	overall	51.885	25.742	2.077	100	$NT = 13696$
	between		24.124	7.035	100	$N = 214$
	within		9.13	5.686	86.947	$T = 64$
GDP Per Capita	overall	12313.794	19265.735	122.679	224582.45	$NT = 11040$
	between		20931.599	302.32	167187.157	$N = 210$
	within		7321.503	-38411.723	113333.003	$T = 52.571$
GDP Per Cap. Growth	overall	1.892	6.495	-64.424	140.491	$NT = 10970$
	between		1.689	-7.984	7.91	$N = 212$
	within		6.317	-66.348	134.472	$T = 51.745$
Post Cold War	overall	0.538	0.499	0	1	$NT = 13975$
	between		0	0.538	0.538	$N = 215$
	within		0.499	0	1	$T = 65$
POLITY	overall	5.551	3.708	0	10	$NT = 8279$
	between		2.985	0	10	$N = 165$
	within		2.229	-1.431	12.319	$T = 50.176$
POLITY Squared	overall	44.569	40.241	0	100	$NT = 8279$
	between		33.233	0	100	$N = 165$
	within		22.91	-30.423	116.513	$T = 50.176$

```
> 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)        0.56660    0.51162    1.11  0.26809
log(LandArea)     -0.10182    0.03053   -3.34  0.00085 ***
log(PopMillions)   0.69431    0.03612   19.22  < 2e-16 ***
UrbanPopulation    0.02205    0.00320    6.89  5.6e-12 ***
log(GDPPerCapita) -0.60751    0.05859  -10.37  < 2e-16 ***
GDPPerCapGrowth   -0.03470    0.00604   -5.74  9.5e-09 ***
PostColdWar       -0.47957    0.08377   -5.73  1.0e-08 ***
POLITY             0.71744    0.06087   11.79  < 2e-16 ***
POLITYSquared     -0.06747    0.00569  -11.86  < 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: 6019.6  on 7366  degrees of freedom
Residual deviance: 4852.8  on 7358  degrees of freedom
  (6670 observations deleted due to missingness)
AIC: 4871

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)     -12.00531    7.65864   -1.57     0.117
log(PopMillions)    0.58488    0.27919    2.09     0.036 *
UrbanPopulation     0.01672    0.01191    1.40     0.160
log(GDPPerCapita)  -0.37336    0.16192   -2.31     0.021 *
GDPPerCapGrowth    -0.05083    0.00792   -6.42   1.4e-10 ***
PostColdWar        -0.17934    0.17435   -1.03     0.304
POLITY              0.69481    0.09203    7.55   4.4e-14 ***
POLITYSquared      -0.07283    0.00872   -8.35   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

residual deviance= 2929,
null deviance= 4551,
n= 4164, N= 86

( 6670 observation(s) deleted due to missingness )
( 3203 observation(s) deleted due to perfect classification )

Number of Fisher Scoring Iterations: 6

Average individual fixed effect= 149.2
```

Alternative Fixed Effects (using `feglm`)

```
> FELogit2<-feglm(CivilWar~log(LandArea)+log(PopMillions)+UrbanPopulation+log(GDPPerCapita)+
+              GDPPerCapGrowth+PostColdWar+POLITY+POLITYSquared|ISO3,data=DF,family="binomial")
NOTES: 6,670 observations removed because of NA values (LHS: 4,985, RHS: 6,670).
       75 fixed-effects (3,203 observations) removed because of only 0 (or only 1) outcomes.

> summary(FELogit2)
GLM estimation, family = binomial, Dep. Var.: CivilWar
Observations: 4,164
Fixed-effects: ISO3: 86
Standard-errors: Clustered (ISO3)
                  Estimate Std. Error z value    Pr(>|z|)
log(LandArea)     -12.00530    8.63479 -1.3903 0.164425407
log(PopMillions)    0.58488    0.74592  0.7841 0.432978449
UrbanPopulation     0.01672    0.03484  0.4799 0.631318817
log(GDPPerCapita)  -0.37336    0.38199 -0.9774 0.328369632
GDPPerCapGrowth    -0.05083    0.01193 -4.2588 0.000020557 ***
PostColdWar        -0.17934    0.46497 -0.3857 0.699717705
POLITY              0.69481    0.23866  2.9112 0.003599910 **
POLITYSquared      -0.07283    0.02361 -3.0852 0.002033893 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Log-Likelihood: -1,464.3   Adj. Pseudo R2: 0.315653
           BIC:  3,712.1     Squared Cor.: 0.397371
```

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

> summary(RELogit)
--------------------------------------------
Maximum Likelihood estimation
BFGS maximization, 86 iterations
Return code 0: successful convergence
Log-Likelihood: -1677
10  free parameters
Estimates:
                  Estimate Std. error t value Pr(> t)
(Intercept)       -0.98257    0.91423   -1.07  0.2825
log(LandArea)     -0.00817    0.05876   -0.14  0.8895
log(PopMillions)   0.74765    0.05995   12.47 < 2e-16 ***
UrbanPopulation    0.01570    0.00519    3.03  0.0025 **
log(GDPPerCapita) -0.65126    0.09661   -6.74 1.6e-11 ***
GDPPerCapGrowth   -0.04889    0.00761   -6.42 1.3e-10 ***
PostColdWar       -0.10801    0.11708   -0.92  0.3562
POLITY             0.67445    0.07998    8.43 < 2e-16 ***
POLITYSquared     -0.07389    0.00771   -9.58 < 2e-16 ***
sigma              2.10909    0.09994   21.10 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
--------------------------------------------
```

Nice Table...

Models of Civil War

	Logit	FE Logit	FEs+Robust	RE Logit
Intercept	0.57			−0.98
	(0.51)			(0.91)
ln(Land Area)	-0.10^*	−12.01	−12.01	−0.01
	(0.03)	(7.66)	(8.63)	(0.06)
ln(Population)	0.69^*	0.58^*	0.58	0.75^*
	(0.04)	(0.28)	(0.75)	(0.06)
Urban Population	0.02^*	0.02	0.02	0.02^*
	(0.00)	(0.01)	(0.03)	(0.01)
ln(GDP Per Capita)	-0.61^*	-0.37^*	−0.37	-0.65^*
	(0.06)	(0.16)	(0.38)	(0.10)
GDP Growth	-0.03^*	-0.05^*	-0.05^*	-0.05^*
	(0.01)	(0.01)	(0.01)	(0.01)
Post-Cold War	-0.48^*	−0.18	−0.18	−0.11
	(0.08)	(0.17)	(0.46)	(0.12)
POLITY	0.72^*	0.69^*	0.69^*	0.67^*
	(0.06)	(0.09)	(0.24)	(0.08)
POLITY Squared	-0.07^*	-0.07^*	-0.07^*	-0.07^*
	(0.01)	(0.01)	(0.02)	(0.01)
Estimated Sigma				2.11^*
				(0.10)
AIC	4870.76			3374.17
BIC	4932.90			
Log Likelihood	−2426.38	−1464.34	−1464.34	−1677.09
Deviance	4852.76	2928.68	2928.68	
Num. obs.	7367	4164	4164	
Num. groups: ISO3			86	
Pseudo R^2			0.32	

$^*p < 0.05$

Models For Event Counts

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

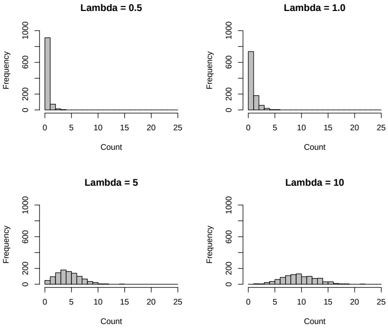
Another motivation: For M independent Bernoulli trials with (sufficiently small) probability of success π and where $M\pi \equiv \lambda > 0$,

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

Poisson: Characteristics

Characteristics:

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



Suppose

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

then

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

with likelihood:

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

and log-likelihood:

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

Event Counts: Unit Effects

The Poisson model:

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

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

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

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

<u>Fixed-Effects Poisson</u>:

- ...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:

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

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

Panel Models: Software

R:

- Tobi$T$ = `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:

- Tobi$T$ = `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)       -1.61539    0.71010   -2.27       0.02291 *
log(LandArea)      0.02638    0.04571    0.58       0.56389
log(PopMillions)   0.43990    0.04481    9.82       < 2e-16 ***
UrbanPopulation    0.00954    0.00453    2.10       0.03536 *
log(GDPPerCapita) -0.47822    0.07684   -6.22 0.00000000049 ***
GDPPerCapGrowth   -0.03443    0.00632   -5.45 0.00000005006 ***
PostColdWar        0.23592    0.11884    1.99       0.04711 *
POLITY             0.31549    0.08256    3.82       0.00013 ***
POLITYSquared     -0.03473    0.00787   -4.41 0.00001016679 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2469.6  on 7366  degrees of freedom
Residual deviance: 2026.7  on 7358  degrees of freedom
  (6670 observations deleted due to missingness)
AIC: 2801

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, 4 iterations
Return code 8: successive function values within relative tolerance limit (reltol)
Log-Likelihood: -1051
8  free parameters
Estimates:
                   Estimate Std. error t value   Pr(> t)
log(LandArea)     -13.53205    6.31007   -2.14    0.0320 *
log(PopMillions)    0.68787    0.30652    2.24    0.0248 *
UrbanPopulation    -0.03983    0.01274   -3.13    0.0018 **
log(GDPPerCapita)  -0.12016    0.13976   -0.86    0.3899
GDPPerCapGrowth    -0.03133    0.00651   -4.81 0.0000015 ***
PostColdWar         0.40306    0.18992    2.12    0.0338 *
POLITY              0.49060    0.10757    4.56 0.0000051 ***
POLITYSquared      -0.05092    0.01048   -4.86 0.0000012 ***
---
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|ISO3,data=DF,family="poisson")

NOTES: 6,670 observations removed because of NA values (LHS: 4,643, RHS: 6,670).
       63 fixed-effects (2,596 observations) removed because of only 0 outcomes.

> summary(FEPoisson2,cluster="ISO3")
GLM estimation, family = poisson, Dep. Var.: OnsetCount
Observations: 4,771
Fixed-effects: ISO3: 98
Standard-errors: Clustered (ISO3)
                   Estimate Std. Error z value      Pr(>|z|)
log(LandArea)     -13.53207   4.140778 -3.2680 0.00108309724 **
log(PopMillions)    0.68788   0.326191  2.1088 0.03496098328 *
UrbanPopulation    -0.03984   0.017334 -2.2981 0.02155734309 *
log(GDPPerCapita)  -0.12016   0.157138 -0.7647 0.44447644897
GDPPerCapGrowth    -0.03133   0.005984 -5.2364 0.00000016372 ***
PostColdWar         0.40306   0.291516  1.3826 0.16677886838
POLITY              0.49060   0.111584  4.3967 0.00001098924 ***
POLITYSquared      -0.05092   0.011461 -4.4432 0.00000886244 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Log-Likelihood: -1,194.2   Adj. Pseudo R2: 0.095584
           BIC:  3,286.3     Squared Cor.: 0.165799
```

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: -1339
10  free parameters
Estimates:
                  Estimate Std. error t value     Pr(> t)
(Intercept)      -2.744822   1.028803   -2.67      0.0076 **
log(LandArea)    -0.000294   0.071785    0.00      0.9967
log(PopMillions)  0.450122   0.078710    5.72 0.000000011 ***
UrbanPopulation  -0.003426   0.006232   -0.55      0.5825
log(GDPPerCapita) -0.236302   0.099028   -2.39      0.0170 *
GDPPerCapGrowth  -0.032882   0.006448   -5.10 0.000000341 ***
PostColdWar       0.263971   0.126335    2.09      0.0367 *
POLITY            0.478502   0.095197    5.03 0.000000500 ***
POLITYSquared    -0.052547   0.009191   -5.72 0.000000011 ***
sigma             1.658763   0.391527    4.24 0.000022688 ***
---
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|ISO3,data=DF)

NOTES: 6,670 observations removed because of NA values (LHS: 4,643, RHS: 6,670).
       63 fixed-effects (2,596 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.

> summary(FENegBin2)
ML estimation, family = Negative Binomial, Dep. Var.: OnsetCount
Observations: 4,771
Fixed-effects: ISO3: 98
Standard-errors: NA (not-available)
                   Estimate Std. Error z value Pr(>|z|)
log(LandArea)     -13.53208         NA      NA       NA
log(PopMillions)    0.68788         NA      NA       NA
UrbanPopulation    -0.03984         NA      NA       NA
log(GDPPerCapita)  -0.12016         NA      NA       NA
GDPPerCapGrowth    -0.03133         NA      NA       NA
PostColdWar         0.40304         NA      NA       NA
POLITY              0.49061         NA      NA       NA
POLITYSquared      -0.05092         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Over-dispersion parameter: theta = 10000 (theta >> 0, no sign of overdispersion, you
  may consider a Poisson model)
Log-Likelihood: -1,194.2   Adj. Pseudo R2: 0.086372
           BIC:  3,286.3     Squared Cor.: 0.165798
```

Table!

Models of Civil War Onset Counts

	Poisson	FE Poisson	RE Poisson
Intercept	-1.62^*		-2.74^*
	(0.71)		(1.03)
ln(Land Area)	0.03	-13.53^*	-0.00
	(0.05)	(6.31)	(0.07)
ln(Population)	0.44^*	0.69^*	0.45^*
	(0.04)	(0.31)	(0.08)
Urban Population	0.01^*	-0.04^*	-0.00
	(0.00)	(0.01)	(0.01)
ln(GDP Per Capita)	-0.48^*	-0.12	-0.24^*
	(0.08)	(0.14)	(0.10)
GDP Growth	-0.03^*	-0.03^*	-0.03^*
	(0.01)	(0.01)	(0.01)
Post-Cold War	0.24^*	0.40^*	0.26^*
	(0.12)	(0.19)	(0.13)
POLITY	0.32^*	0.49^*	0.48^*
	(0.08)	(0.11)	(0.10)
POLITY Squared	-0.03^*	-0.05^*	-0.05^*
	(0.01)	(0.01)	(0.01)
Estimated Sigma			1.66^*
			(0.39)
AIC	2800.92	2118.74	2697.27
BIC	2863.07		
Log Likelihood	-1391.46	-1051.37	-1338.63
Deviance	2026.74		
Num. obs.	7367		

$^*p < 0.05$

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

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

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

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

Define:

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

$\rightarrow$ "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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Score equations:

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

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):

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

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

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

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

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

- $\hat{\Sigma}_{\text{Model}}$
  - Requires that $\mathbf{R}_i(\boldsymbol{\alpha})$ be "correct" for consistency.
  - Is slightly more efficient than $\hat{\boldsymbol{\Sigma}}_{\text{Robust}}$ if so.

- $\hat{\Sigma}_{\text{Robust}}$
  - Is consistent *even if* $\mathbf{R}_i(\boldsymbol{\alpha})$ *is misspecified*.
  - Is slightly less efficient than $\hat{\boldsymbol{\Sigma}}_{\text{Model}}$ if $\mathbf{R}_i(\boldsymbol{\alpha})$ is correct.

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

Summary

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 βs,
- 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, $\hat{\boldsymbol{\beta}}_M \approx \frac{\hat{\boldsymbol{\beta}}_C}{\sqrt{1+0.35\sigma_\eta^2}}$, 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 $\mathbf{R}_i(\boldsymbol{\alpha})$

- Has been called "more art than science."
- Pointers:
  - Choose based on *substance* of the problem.
  - Remember that $\mathbf{R}_i(\boldsymbol{\alpha})$ is conditional on $\mathbf{X}$, $\hat{\boldsymbol{\beta}}$.
  - 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)

GEEs: Software Tips

- 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=ISO3,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 = ISO3,
    corstr = "independence")

 Coefficients:
                  Estimate Std.err  Wald   Pr(>|W|)
(Intercept)         0.5666  1.9536  0.08    0.77179
log(LandArea)      -0.1018  0.1215  0.70    0.40195
log(PopMillions)    0.6943  0.1388 25.02 0.00000057 ***
UrbanPopulation     0.0220  0.0135  2.66    0.10308
log(GDPPerCapita)  -0.6075  0.2274  7.14    0.00755 **
GDPPerCapGrowth    -0.0347  0.0120  8.42    0.00371 **
PostColdWar        -0.4796  0.2584  3.45    0.06343 .
POLITY              0.7174  0.2131 11.33    0.00076 ***
POLITYSquared      -0.0675  0.0193 12.19    0.00048 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = independence
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)    0.845   0.433
Number of clusters:   161  Maximum cluster size: 57
```

GEE: Exchangeable

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

 Coefficients:
                  Estimate  Std.err  Wald Pr(>|W|)
(Intercept)       -3.80033  2.71926  1.95   0.1622
log(LandArea)      0.19709  0.23788  0.69   0.4074
log(PopMillions)   0.48603  0.19507  6.21   0.0127 *
UrbanPopulation    0.00166  0.01215  0.02   0.8913
log(GDPPerCapita) -0.28648  0.20287  1.99   0.1579
GDPPerCapGrowth   -0.03519  0.00871 16.33 0.000053 ***
PostColdWar       -0.06790  0.23875  0.08   0.7761
POLITY             0.50948  0.16628  9.39   0.0022 **
POLITYSquared     -0.05256  0.01650 10.15   0.0014 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = exchangeable
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)    0.935    1.97
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.504   0.674
Number of clusters:   161  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)       -1.55942  9.52623 0.03    0.870
log(LandArea)      0.08399  0.67323 0.02    0.901
log(PopMillions)  -0.44625  0.79798 0.31    0.576
UrbanPopulation   -0.04251  0.10101 0.18    0.674
log(GDPPerCapita)  0.01190  0.81715 0.00    0.988
GDPPerCapGrowth   -0.01758  0.00835 4.44    0.035 *
PostColdWar        0.30524  0.48295 0.40    0.527
POLITY             0.21724  0.29136 0.56    0.456
POLITYSquared     -0.02316  0.03006 0.59    0.441
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = ar1
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)     6.85     316
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.971    2.16
Number of clusters:   161  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=ISO3,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 = ISO3,
    corstr = "unstructured")

 Coefficients:
                  Estimate Std.err  Wald Pr(>|W|)
(Intercept)        -1.9652  3.1213  0.40  0.52896
log(LandArea)       0.1374  0.1911  0.52  0.47201
log(PopMillions)    0.8891  0.2451 13.16  0.00029 ***
UrbanPopulation     0.0347  0.0168  4.24  0.03940 *
log(GDPPerCapita)  -0.8357  0.2994  7.79  0.00525 **
GDPPerCapGrowth    -0.0160  0.0309  0.27  0.60496
POLITY              0.5023  0.4096  1.50  0.22005
POLITYSquared      -0.0581  0.0360  2.61  0.10644
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation structure = unstructured
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)    0.667   0.789
  Link = identity

Estimated Correlation Parameters:
          Estimate Std.err
alpha.1:2    0.384   0.477
alpha.1:3    0.405   0.502
alpha.1:4    0.345   0.431
alpha.1:5    0.334   0.415
alpha.2:3    0.714   0.822
alpha.2:4    0.279   0.352
alpha.2:5    0.496   0.573
alpha.3:4    0.397   0.501
alpha.3:5    0.738   0.857
alpha.4:5    0.435   0.542
Number of clusters:   159  Maximum cluster size: 5
```

GEE Model Comparison

Table 1: GEE Models of Civil War Onset

	Independence	Exchangeable	AR(1)	Unstructured (2013-17)
(Intercept)	0.567	−3.800	−1.559	−1.965
	(1.954)	(2.719)	(9.526)	(3.121)
ln(Land Area)	−0.102	0.197	0.084	0.137
	(0.121)	(0.238)	(0.673)	(0.191)
ln(Population)	0.694***	0.486*	−0.446	0.889***
	(0.139)	(0.195)	(0.798)	(0.245)
Urban Population	0.022	0.002	−0.043	0.035*
	(0.014)	(0.012)	(0.101)	(0.017)
ln(GDP Per Capita)	−0.608**	−0.286	0.012	−0.836**
	(0.227)	(0.203)	(0.817)	(0.299)
GDP Growth	−0.035**	−0.035***	−0.018*	−0.016
	(0.012)	(0.009)	(0.008)	(0.031)
Post-Cold War	−0.480+	−0.068	0.305	
	(0.258)	(0.239)	(0.483)	
POLITY	0.717***	0.509**	0.217	0.502
	(0.213)	(0.166)	(0.291)	(0.410)
POLITY Squared	−0.067***	−0.053**	−0.023	−0.058
	(0.019)	(0.016)	(0.030)	(0.036)
Num.Obs.	7367	7367	7367	791

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

GEEs: Wrap-Up

<u>GEEs are</u>:

- 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:

$$
\begin{aligned}
Y_i &= \text{the duration until the event occurs,} \\
Z_i &= \text{the duration until the observation is ``censored''} \\
T_i &= \min\{Y_i, Z_i\}, \\
C_i &= \text{0 if observation } i \text{ is } \textit{censored}\text{, 1 if it is not.}
\end{aligned}
$$

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:

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

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:

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

$\longrightarrow$ 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...

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

Temporal Issues in Grouped-Data Models

"Time dummies":

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

$\longrightarrow$ 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.)