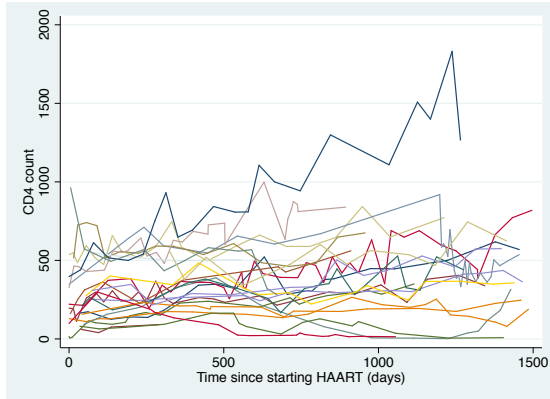


# Longitudinal Data

## Fall 2015



### Chapter 7, part 3

## **Mixed, Random Effects, Random Coefficients, Multilevel, ...Models**

### Instructors

Nick Jewell ([jewell@berkeley.edu](mailto:jewell@berkeley.edu))



### GSI

Robin Mejia ([mejia@nasw.org](mailto:mejia@nasw.org))

# General Mixed Model for Logistic Regression - Theoretically

- Similar to linear model, can be generally described as a model with the logit of the probability of the outcome conditional on covariates and the specific unit (and possibly sub-unit).
- Consists of parameters  $(\beta)$ \*covariates( $X_{ijk..}$ ) and random variables( $U_i$ )\*covariates( $Z_{ijk..}$ ), where usually  $Z_{ijk..}$  is a subset of  $X_{ijk..}$ .

$$\log it \left[ P(Y_{ijk...} = 1 \mid \vec{X}_{ijk..}, \vec{U}_i, \vec{Z}_{ijk...}) \right] = \vec{X}_{ijk..}^T \vec{\beta} + \vec{Z}_{ijk...}^T \vec{U}_i$$

# General Mixed Model for Logistic Regression - Theoretically

- Assumption is typically  $U_i \sim \text{MVN}(0, \Sigma)$ .
- Solve using MLE -

$$L_{\beta, \Sigma}(\vec{Y}_i \mid \mathbf{X}_i) = \int_{\vec{U}_i} P(Y_i \mid \mathbf{X}_i, \mathbf{Z}_i, \vec{U}_i) dP(\vec{U}_i)$$

# General Mixed Model for Logistic Regression - Practically

Like linear mixed models, one can use these models to:

1. Introduce complicated correlation structures to account for repeated measures structures.
  2. Model source of variation at different hierarchical levels.
  3. Get estimates of *within unit* associations as opposed as population average associations (from GEE models).
  4. Predict random effects.
- Use xtmelogit in STATA.

# Random Effects Model for Teenage Sex and Drug-Use

$$\log it[P(Y_{ij} = 1 | \beta_{0i}, X_{ij} = x_{ij})] = \log \left( \frac{P(Y_{ij} = 1 | \beta_{0i}, X_{ij} = x_{ij})}{P(Y_{ij} = 0 | \beta_{0i}, X_{ij} = x_{ij})} \right) = \beta_0^* + \beta_{0i} + \beta_1^* x_{ij}$$

- Assume that the repeated observations for the  $i$ th teenager are independent of one another given  $\beta_{i0}$  and  $X_{ij}$ .
- Must assume parametric distribution for the  $\beta_{i0}$ , usually  $\beta_{i0} \sim N(0, \tau^2)$ .
- $\exp(\beta_1^*)$  is odds ratio for having sex infection when subject  $i$  reports drug-use relative to when same subject does not report drug-use.

# Teenage Sex and Drug-Use

## Number of observations per teen

```
tab cattot if cnt==1
```

Total # Obs			
per teen	Freq.	Percent	Cum.
-----+-----			
1-10	35	31.82	31.82
11-20	22	20.00	51.82
21-30	48	43.64	95.45
>31	5	4.55	100.00
-----+-----			
Total	110	100.00	

## Proportion of days with Sexual Activity by Number of observations total on a teen

Total #			
Obs per	sx24hrs		
teen	no	yes	Total
-----+-----			
1-10	98	68	166
	59.04	40.96	100.00
-----+-----			
11-20	225	100	325
	69.23	30.77	100.00
-----+-----			
21-30	928	291	1,219
	76.13	23.87	100.00
-----+-----			
>31	156	43	199
	78.39	21.61	100.00
-----+-----			
Total	1,407	502	1,909
	73.70	26.30	100.00

# Proportion of days with Drug/Alcohol use by Number of observations total on a teen

Total #			
Obs per	drgalcoh		
teen	no	yes	Total
-----+-----+-----			
1-10	88	68	156
	56.41	43.59	100.00
-----+-----+-----			
11-20	205	112	317
	64.67	35.33	100.00
-----+-----+-----			
21-30	840	256	1,096
	76.64	23.36	100.00
-----+-----+-----			
>31	118	21	139
	84.89	15.11	100.00
-----+-----+-----			
Total	1,251	457	1,708
	73.24	26.76	100.00

. cs sx24hrs drgalcoh, or

	drgalcoh			
	Exposed	Unexposed		Total
-----+-----+-----				
Cases	171	320		491
Noncases	286	931		1217
-----+-----+-----				
Total	457	1251		1708
Risk	.3741794	.2557954		.2874707
	Point estimate			[95% Conf. Interval]
Risk difference	.1183841			.0678575 .1689107
Risk ratio	1.462808			1.256995 1.702318
Attr. frac. ex.	.3163832			.2044521 .4125659
Attr. frac. pop	.1101864			
Odds ratio	1.739521			1.385047 2.184751 (Cornfield)
-----+-----				
chi2(1) = 22.90 Pr>chi2 = 0.0000				



## Random effects for teenage sex vs drug use

```
. xtmelogit sx24hrs drgalcoh || eid:, stddeviations
```

sx24hrs	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
drgalcoh	.3882629	.1552394	2.50	0.012	.0839993	.6925264
_cons	-1.08007	.1537975	-7.02	0.000	-1.381507	-.778632

Random-effects Parameters		Estimate	Std. Err.	[95% Conf. Interval]	
eid: Identity					
	sd(_cons)	1.274216	.1424908	1.023426	1.586462

LR test vs. logistic regression:  $\chi^2(01) = 183.36$  Prob>=  $\chi^2 = 0.0000$

```
. lincom drgalcoh, or
```

sx24hrs	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
(1)	<b>1.474417</b>	.2288876	2.50	<b>0.012</b>	1.087628	1.998759

# Random effects for teenage sex vs drug use

```
. estat ic
```

Akaike's information criterion and Bayesian information criterion

-----							
Model		Obs	ll (null)	ll (model)	df	AIC	BIC
-----+-----							
.		1708	.	-921.7965	3	1849.593	1865.922
-----							

Note: N=Obs used in calculating BIC; see [R] BIC note

```
. estat icc
```

Residual intraclass correlation

-----				
Level		ICC	Std. Err.	[95% Conf. Interval]
-----+-----				
eid		.3304423	.0494831	.2414885 .4334388
-----				

# Random effects for teenage sex vs drug use

- So how do they calculate the ICC?
  - estat command gave ICC = 0.3304423
  - results gave  $\sigma^2_u = 1.274216^2 = 1.623626$
  - $ICC = \sigma^2_u / (\sigma^2_u + \sigma^2_e) \rightarrow \sigma^2_e = \mathbf{3.289867}$
  - Where does this number come from?
  - Stata assumes individual errors come from the logistic distribution, mean 0, variance  $\pi^2 / 3$  (also parameterized as location 0, scale 1)

# Re-do to get Robust SE's with Clustered Bootstrap

```
set seed 123456
bootstrap, cluster(eid) idcluster(newid) group(eid) reps(50): meqrlogit sx24hrs drgalcoh ||
newid:, stddeviations
(running meqrlogit on estimation sample)
```

Bootstrap replications (50)

-----+----- 1 -----+----- 2 -----+----- 3 -----+----- 4 -----+----- 5

..... 50

Mixed-effects logistic regression                      Number of obs            =        1708  
Group variable: newid                                    Number of groups        =        109

Obs per group: min =        1  
                  avg =       15.7  
                  max =       33

Integration points =       7                            Wald chi2(1)            =        6.07  
Log likelihood = -921.79647                            Prob > chi2            =        0.0138

(Replications based on 109 clusters in eid)

		Observed	Bootstrap	Normal-based			
sx24hrs		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
drgalcoh		.3882629	.1576392	2.46	0.014	.0792958	.6972299
_cons		-1.08007	.1580662	-6.83	0.000	-1.389874	-.7702655

Random-effects Parameters		Observed	Bootstrap	Normal-based	
		Estimate	Std. Err.	[95% Conf. Interval]	
newid: Identity					
	var(_cons)	1.623626	.4010462	1.000538	2.634743

LR test vs. logistic regression: chibar2(01) =    183.36 Prob>=chibar2 = 0.0000

# Random *coefficients* model for teenage sex vs drug use

$$\log it[P(Y_{ij} = 1 | \beta_{0i}, \beta_{1i}, X_{ij} = x_{ij})] = \log \left( \frac{P(Y_{ij} = 1 | \beta_{0i}, \beta_{1i}, X_{ij} = x_{ij})}{1 - P(Y_{ij} = 1 | \beta_{0i}, \beta_{1i}, X_{ij} = x_{ij})} \right) = (\beta_0 + \beta_{0i}) + (\beta_1 + \beta_{1i})x_{ij}$$

```
. meqrlogit sx24hrs drgalcoh || eid: drgalcoh, stddeviations cov(unstruct)
```

[a bunch of information removed from here]

sx24hrs	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
drgalcoh	.3287796	.1913386	1.72	0.086	-.0462372	.7037965
_cons	-1.077791	.1437309	-7.50	0.000	-1.359498	-.7960831

## Random *coefficients* model for teenage sex vs drug use

$$\log it[P(Y_{ij} = 1 | \beta_{0i}, \beta_{1i}, X_{ij} = x_{ij})] = \log \left( \frac{P(Y_{ij} = 1 | \beta_{0i}, \beta_{1i}, X_{ij} = x_{ij})}{1 - P(Y_{ij} = 1 | \beta_{0i}, \beta_{1i}, X_{ij} = x_{ij})} \right) = (\beta_0 + \beta_{0i}) + (\beta_1 + \beta_{1i})x_{ij}$$

```

Random-effects Parameters | Estimate Std. Err. [95% Conf. Interval]
-----+-----
eid: Unstructured |
      sd(drgalcoh) |   .7279197   .2996808   .3248241   1.631243
      sd(_cons) |   1.16517   .1544805   .8985357   1.510926
      corr(drgalcoh,_cons) |   .4527734   .3999896   -.4604965   .900396
-----+-----
LR test vs. logistic regression:   chi2(3) =   187.82   Prob > chi2 = 0.0000
. lincom drgalcoh, or
-----+-----
      sx24hrs | Odds Ratio Std. Err.      z    P>|z|      [95% Conf. Interval]
-----+-----
      (1) |   1.389272   .2658214   1.72   0.086   .9548154   2.021412

```

estat ic

Akaike's information criterion and Bayesian information criterion

-----						
Model		Obs	ll(null)	ll(model)	df	AIC            BIC
-----+-----						
.		1708	.	-919.5649	5	1849.13    1876.345
-----						

Note: N=Obs used in calculating BIC; see [R] BIC note

. estat icc

Conditional intraclass correlation

-----				
Level		ICC	Std. Err.	[95% Conf. Interval]
-----+-----				
eid		.2921193	.0548321	.1970516    .4096525
-----				

Note: ICC is conditional on zero values of random-effects covariates.

Model	Parameter	Estimate	SE(Naïve)
Simple Random Effects	$\beta_0$	-1.08	0.15
	$\beta_1$	0.39	0.15
	OR	1.47	0.23
	SD( $\beta_{0i}$ )	1.27	0.14
	ICC	0.33	0.049
Fit Statistic		AIC	1849.6
Random Coefficients	$\beta_0$	-1.08	0.14
	$\beta_1$	0.33	0.19
	OR	1.38	0.27
	SD( $\beta_{0i}$ )	1.16	0.15
	SD( $\beta_{1i}$ )	0.73	0.30
	ICC	0.29	0.055
	Cor( $\beta_{0i}, \beta_{2i}$ )	0.45	0.40
Fit Statistic		AIC	1849.1



# Range of Impacts (defined by odds ratio) of drug/alcohol use on sexual activity

$$\log it[P(Y_{ij} = 1 | \beta_{0i}, \beta_{1i}, X_{ij} = x_{ij})] = \log \left( \frac{P(Y_{ij} = 1 | \beta_{0i}, \beta_{1i}, X_{ij} = x_{ij})}{1 - P(Y_{ij} = 1 | \beta_{0i}, \beta_{1i}, X_{ij} = x_{ij})} \right) = (\beta_0 + \beta_{0i}) + (\beta_1 + \beta_{1i})x_{ij}$$

- The estimated IQR of the odds ratios in pop<sup>n</sup>

is  $(\exp\{\hat{\beta}_1 - Z_{0.75} * \hat{\sigma}_{\beta_{1i}}\}, \exp\{\hat{\beta}_1 + Z_{0.75} * \hat{\sigma}_{\beta_{1i}}\}) =$   
 $(\exp\{0.33 - 0.67 * 0.73\}, \exp\{0.33 + 0.67 * 0.73\})$

```
. display "(" exp(.3287796-invnormal(0.75)*.7279197)      " , "
               exp(.3287796+invnormal(0.75)*.7279197)  " )"
.      (.85027557 , 2.2699413 )
```

**This interval contains 1.**

# Random coefficients model adjusting for *history of sexual activity*

Does using a transition model in this context  
give us more information?

$$\log it[P(Y_{ij} = 1 | \beta_{0i}, \beta_{1i}, X_{ij} = x_{ij}, \bar{Y}_{i,j-1})] = (\beta_0 + \beta_{0i}) + (\beta_1 + \beta_{1i})x_{ij} + \beta_2 \bar{Y}_{i,j-1}$$

```
*** Creating "cumulative average"
```

```
sort eid time
```

```
capture drop cumsum cumprop cumlag
```

```
gen cumsum = 0
```

```
replace cumsum = sx24hrs if ct==1
```

```
by eid: replace cumsum = cumsum[_n-1]+sx24hrs if ct > 1
```

```
gen cumprop = cumsum/ct
```

```
by eid: gen cumlag = cumprop[_n-1] if ct > 1
```

```
replace cumlag = 0 if ct==1
```

## Random coefficients model adjusting for *history of sexual activity*

$$\log it[P(Y_{ij} = 1 | \beta_{0i}, \beta_{1i}, X_{ij} = x_{ij}, \bar{Y}_{i,j-1})] = (\beta_0 + \beta_{0i}) + (\beta_1 + \beta_{1i})x_{ij} + \beta_2 \bar{Y}_{i,j-1}$$

```
megrlogit sx24hrs drgalcoh cumlag || eid: drgalcoh, stddeviations cov(unstruct)
```

sx24hrs	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
drgalcoh	.3158968	.2025083	1.56	0.119	-.0810122 .7128059
<b>cumlag</b>	-.6029536	.5978829	-1.01	0.313	-1.774783 .5688754
_cons	-.8357194	.2291452	-3.65	0.000	-1.284836 -.386603

lincom drgalcoh, or

sx24hrs	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
(1)	1.371489	.2777379	1.56	0.119	.9221824 2.039706

# Compare random effects models

Did adding history of sexual activity to the model predicting sex in the last 24 hours from drug/alcohol use make a statistical difference?

Model	OR	95% CI
Random slope, without cumlag	1.389	(.955, 2.021)
Random slope, with cumlag	1.371	(.922, 2.040)

```
lrtest A B
```

```
Likelihood-ratio test
```

```
LR chi2(1) = 0.61
```

```
(Assumption: B nested in A)
```

```
Prob > chi2 = 0.4333
```

# Random Effects Model for Diarrhea Study in Children

$$\log\left(\frac{P(Y_{ijk} = 1)}{1 - P(Y_{ijk} = 1)}\right) = \beta_0 + \beta_{0i} + \beta_{0ij}$$

- Measurements made on children (k) within households (j) within villages (i).
- Question of interest: Are households or villages the greatest sources of variation?
  - households –  $var(\beta_{0ij})$  or
  - villages –  $var(\beta_{0i})$
- Assumes children in same household have same probability of diarrhea.
  - Use meqrlogit in STATA

# Random Effects Model for Diarrhea Study in Children

```
xtmelogit diarrhea || vilid: || hhid:, intpoints(5)
```

Mixed-effects logistic regression      Number of obs      =      30371

Group	Variable	No. of Groups	Observations per Group			Integration Points
			Minimum	Average	Maximum	
	vilid	12	610	2530.9	5452	5
	hhid	95	20	319.7	672	5

```
Log likelihood = -8614.8567      Wald chi2(0)      =      .
                                Prob > chi2      =      .
```

diarrhea	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_cons	-2.482402	.082076	-30.25	0.000	-2.643268 -2.321536

# Random Effects Model for Diarrhea Study in Children

## Variance of Random Effects

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
-----+-----				
vilid: Identity <i>estim. SD</i> ( $\beta_{0i}$ )				
sd(_cons)	.0001116	.0211476	4.8e-166	2.6e+157
-----+-----				
hhid: Identity <i>estim. SD</i> ( $\beta_{0ij}$ )				
sd(_cons)	.7594603	.0651984	.641846	.8986268
-----+-----				
LR test vs. logistic regression:	chi2(2) =	828.69	Prob > chi2 = 0.0000	

Cluster correlation coefficient (based on latent response model) – not as simple for logistic mixed models as it is for linear

$$\hat{\rho}_{house} = \frac{\text{var}(\beta_{0ij})}{\text{var}(\beta_{0ij}) + \text{var}(\beta_{0i}) + \pi^2/3} = \frac{0.76^2}{0.76^2 + 0.0001^2 + 3.29} = 0.15$$

$$\hat{\rho}_{village} = \frac{\text{var}(\beta_{0i})}{\text{var}(\beta_{0ij}) + \text{var}(\beta_{0i}) + \pi^2/3} = \frac{0.0001^2}{0.76^2 + 0.0001^2 + 3.29} = 0.00$$

Using conditional logistic regression  
for estimating within unit OR in  
logistic regression models



## Treat Individual as a stratification variable for Teen Sex and Drugs

- For the teen sex and drugs example, we can represent the data on each individual,  $i$ , as a simple 2x2 table:

Sex		yes	no	
D r u g s	yes	$a_i$	$b_i$	
	no	$c_i$	$d_i$	
				$n_i$

- Can get the OR for every subject:  $\hat{OR}_i = \frac{a_i d_i}{b_i c_i}$

- $$\log it[P(Y_{ij} = 1 \mid \beta_{0i}, X_{ij} = x_{ij})] = \log \left( \frac{P(Y_{ij} = 1 \mid \beta_{0i}, X_{ij} = x_{ij})}{P(Y_{ij} = 0 \mid \beta_{0i}, X_{ij} = x_{ij})} \right) = \beta_0^* + \beta_{0i} + \beta_1^* x_{ij}$$

assumes every person has the same OR, so one can average each estimated OR to get the estimate.

# Mantel-Haenszel Average of Stratified OR's

- Then the MH estimate is:

$$\hat{OR}_{MH} = \exp(\hat{\beta}_1^*) = \frac{\sum_{i=1}^m w_i \hat{OR}_i}{\sum_{i=1}^m w_i} = \frac{\sum_{i=1}^m (a_i d_i) / n_i}{\sum_{i=1}^m (b_i c_i) / n_i}$$

- Note, that for any subject who has identical exposure (drug use) or outcomes (sex) for all observations, the OR is undefined and that person does not contribute to the estimate (their 2x2 table are dropped).

# Conditional Logistic Regression

$$\log it[P(Y_{ij} = 1 | \beta_{0i}, X_{ij} = x_{ij})] = \log \left( \frac{P(Y_{ij} = 1 | \beta_{0i}, X_{ij} = x_{ij})}{P(Y_{ij} = 0 | \beta_{0i}, X_{ij} = x_{ij})} \right) = \beta_0^* + \beta_{0i} + \beta_1^* x_{ij}$$

- To illustrate, use the teenage sex and drugs example, assume just two observation for a person, and that one had the outcome ( $Y_{i1}=1$ ) with drugs ( $X_{i1}=1$ ) one observation had neither ( $Y_{i2}=0, X_{i2}=0$ ).
- Then, the conditional likelihood contribution for this observation is:

$$CondLik_i = \frac{P(Y_{i1} = 1 | X_{i1} = 1)P(Y_{i2} = 0 | X_{i2} = 0)}{P(Y_{i1} = 1 | X_{i1} = 1)P(Y_{i2} = 0 | X_{i2} = 0) + P(Y_{i1} = 1 | X_{i1} = 0)P(Y_{i2} = 0 | X_{i2} = 1)}$$

# Conditional Logistic Regression

- After plugging in the model for  $Y_{ij}$  and doing some algebra, one gets:

$$CondLik_i = \frac{1}{1 + \exp(\beta_1^* (X_{i2} - X_{i1}))}$$

- Notice, the individual level intercept (whether random or not) drops out.

# Conditional Logistic Regression

- What it means is that the estimate of the within subject OR no longer depends on assumptions on the distribution of the random effect.
- Can only use this to estimate the association of time-varying covariates.
- Subjects with identical outcomes will be dropped from analysis.
- For those covariates that do not change in a subject, they will not contribute to estimation of the OR for that covariate.

# Conditional Logistic Regression

- More generally, you might want to estimate the within subject OR for several variables simultaneously and/or the OR for a unit change in a continuous variable.
- Can still do so by using the conditional likelihood - a method used to estimate OR's for matched case-control studies.
- The conditional likelihood (in example of a cohort) is the probability of observing that the cases have covariates they have and the controls have their observed covariates, given the distribution of covariates observed over all the repeated measurements.
- To define the likelihood, one normalizes the probability of observing the outcomes conditional on the covariates by the summed probabilities over all possible combinations of covariates and outcomes.

# Teenage Sex and Drug-Use Using M-H summary OR.

```
. cs sx24hrs drgalcoh, by(eid) or
```

	eid	OR	[95% Conf. Interval]	M-H Weight
	-----+-----			
1	.	.	.	0 (Cornfield)
2	0	0	10.56942	.3478261 (Cornfield)
3	.	0	.	0 (Cornfield)
4	.	.	.	0 (Cornfield)
5	.	.	.	0 (Cornfield)
6	1.333333	.2058078	8.53481	.8823529 (Cornfield)
7	.	.	.	0 (Cornfield)
8	.	0	.	0 (Cornfield)
9	1.5	.1778039	12.91562	.5714286 (Cornfield)
10	.	0	.	0 (Cornfield)
105	.	.	.	0 (Cornfield)
106	0	0	.	.6363636 (Cornfield)
107	.8	.1388054	4.9008	1.2 (Cornfield)
108	0	0	.	.125 (Cornfield)
109	.	.	.	0 (Cornfield)
110	.	0	.	0 (Cornfield)
	-----+-----			
MH Combined	1.315498	.9584698	1.805519	
	-----			

# Conditional Logistic Estimate

```
. clogit sx24hrs drgalcoh, or group(eid)
note: multiple positive outcomes within groups encountered.
note: 23 groups (161 obs) dropped due to all positive or
      all negative outcomes.
```

```
Iteration 0:    log likelihood = -664.37829
Iteration 1:    log likelihood = -663.20668
Iteration 2:    log likelihood = -663.20668
```

Conditional (fixed-effects) logistic regression	Number of obs	=	1547
	LR chi2(1)	=	2.93
	Prob > chi2	=	0.0867
Log likelihood = -663.20668	Pseudo R2	=	0.0022

-----						
sx24hrs	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
drgalcoh	1.323141	.2158621	1.72	0.086	.9610325	1.821689
-----						