Longitudinal Data Analysis

GENERALIZED LINEAR MIXED MODELS (GLMMs)

Categorical Response Variables

Q: If we have multivariate categorical data then what models / methods are available?

Examples

- Thymectomy Data (ordinal response scale)
- Madras Symptom Data (binary response)
- Infection Data (Xeropthalmia, Six Cities)
- BSS Attitudes Data (multilevel binary response)
- Seizure Counts (Progabide Data)

Categorical Response Variables

Recall: VPS IC Analysis

◦ Q4SAFE – "We can be sure that the HIV vaccine is safe once we begin phase III testing"

EDA – time cross-sectional

Baseline

ICgroup	q4safe0		
	0	1	RowTotl
0	218	282	500
	0.44	0.56	0.5
1	216	284	500
	0.43	0.57	0.5

EDA – time cross-sectional

Post-Intervention

ICgroup	q4safe6		
	0	1	RowTotl
0	226	274	500
	0.45	0.55	0.5
1	180	320	500
	0.36	0.64	0.5

ICgroup	q4safe12		
	0	1	RowTotl
0	208	292	500
	0.42	0.58	0.5
1	177	323	500
	0.35	0.65	0.5

EDA – transitions

IC Control Group

q4safe0 \ q4safe6	0	1	RowTotl
0	148	70	218
	0.68	0.32	0.44
	0.65	0.26	
1	78	204	282
	0.28	0.72	0.56
	0.35	0.74	
ColTotl	226	274	500
	0.45	0.55	

EDA – transitions

IC Intervention Group

q4safe0 | q4safe6	0	1	RowTotl
0	118	98	216
	0.55	0.45	0.43
	0.66	0.31	
1	62	222	284
	0.22	0.78	0.57
	0.34	0.69	
ColTotl	180	320	500
	0.36	0.64	

Regression Models

Q: Is there an intervention effect? If so what is it?

Q: Does the intervention effect "wane"?

Regression Models:

$$
\begin{aligned}
Y_{ij} &= \text{response at time } j \text{ for subject } i \\
\mu_{ij} &= E(Y_{ij} \mid X_{ij})
\end{aligned}
$$

$$
\begin{aligned}
\text{logit}(\mu_{ij}) &= \beta_0 + \beta_1 \cdot (\text{Tx}) + \\
&\quad \beta_2 \cdot (\text{Time=6}) + \beta_3 \cdot (\text{Time=12}) + \\
&\quad \beta_4 \cdot (\text{Time=6} \cdot \text{Tx}) + \beta_5 \cdot (\text{Time=12} \cdot \text{Tx})
\end{aligned}
$$

Regression Models

Analysis Options:

- Semi-parametric methods (GEE)

⋆ **"Random effects" models**.

- Transition models

Conditional Regression Models

Q: Can we explicitly "account" for subject heterogeneity in the regression model?

Conditional Regression Models:

$$
\begin{aligned}
Y_{ij} &= \text{response at time } j \text{ for subject } i \\
\mu^b_{ij} &= E(Y_{ij} \mid X_{ij}, b_i)
\end{aligned}
$$

$$
\begin{aligned}
\text{logit}(\mu^b_{ij}) = \; & \boxed{b_i} + \beta_0 + \beta_1 \cdot (\text{Tx}) + \\
& \beta_2 \cdot (\text{Time=6}) + \beta_3 \cdot (\text{Time=12}) + \\
& \beta_4 \cdot (\text{Time=6} \cdot \text{Tx}) + \beta_5 \cdot (\text{Time=12} \cdot \text{Tx})
\end{aligned}
$$

Conditional Regression Models

$\star\star\star$ Assume that $[Y_{ij}, Y_{ik} \mid b_i] = [Y_{ij} \mid b_i][Y_{ik} \mid b_i] \quad \Rightarrow$ conditional independence.

Estimation Options:

- Conditional likelihood methods (eliminate)
- Marginal likelihood methods (integrate)

Parameter Interpretation

- The introduction of b_i is useful for modelling the dependence in the data. That is, outcomes taken on the same individual are more likely to be similar due to the shared (unobserved) factor, b_i.

Q: Does this have any impact on the interpretation of the regression parameters?

Within-cluster covariates

- $\beta_2 \cdot (\text{Time}{=}6)$

- $\beta_4 \cdot (\text{Time}{=}6 \cdot \text{Tx})$

Parameter Interpretation

Between-cluster covariates

- $\beta_1 \cdot (\mathsf{Tx})$

- Any additional person-level covariates (age, education)

Model Interpretation

ZEGER, LIANG, and ALBERT (1988)

Consider a single binary covariate X_{ij} that equals 1 if a child's mother is a smoker and 0 otherwise. Let Y_{ij} denote whether child i experienced a respiratory infection during period j

$$\text{logit} E[Y_{ij} \mid X_{ij}] = \beta_0 + \beta_1 X_{ij}$$

Then β_1 is the population average contrast.

Model Interpretation

ZEGER, LIANG, and ALBERT (1988)

If we postulate a random intercept, b_i, (child propensity for infection) then we may consider the model:

$$\text{logit}E[Y_{ij} \mid X_{ij}, b_i] = b_i + \beta_0^* + \beta_1^* X_{ij}$$

Then β_1^* is the subject-specific contrast.

Model Interpretation

NEUHAUS, KALBFLEISCH, and HAUCK (1991)

"Thus β_1^* measures the change in the conditional logit of the probability of response with the covariate X for individuals in each of the underlying risk groups described by b_i." (pg 20)

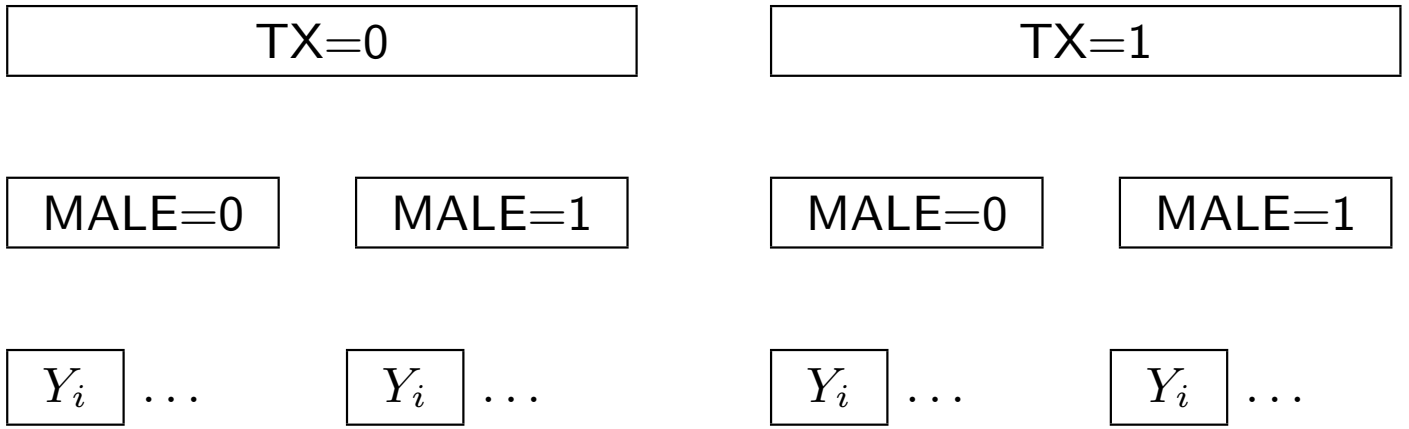
NEUHAUS, KALBFLEISCH, and HAUCK (1991)

"Although the cluster-specific model seems to provide the more unified approach, parameter interpretation in these models is difficult. The cluster-specific model presupposes the existence of latent risk groups indexed by b_i, and parameter interpretation is with reference to these groups. No empirical verification of this statement can be available from the data unless the latent risk groups can be identified. Since each individual is assumed to have her own latent risk b_i, the model almost invites an unjustified causal statement about the change in odds of fluid availability for a given woman who ceases to be nulliparous."

Multilevel Model Example

Scenario 1: (1 level)

Consider the following design:



MARGINAL

$$\text{logit}E[Y_i \mid X_i] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

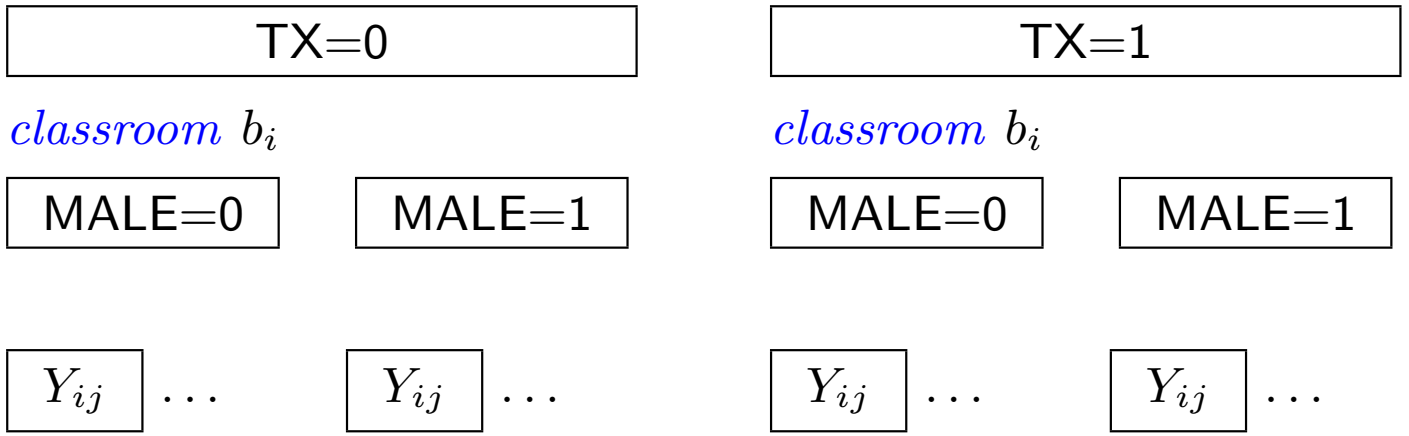
Multilevel Model Example

Scenario 2: (2 level)

Consider the following design:

- subjects clustered in classrooms

Scenario 2



MARGINAL

$$\text{logit}E[Y_{ij} \mid X_{ij}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2ij}$$

CONDITIONAL

$$\text{logit}E[Y_{ij} \mid X_{ij},\ b_i] = b_i + \beta_0^* + \beta_1^* X_{1i} + \beta_2^* X_{2ij}$$

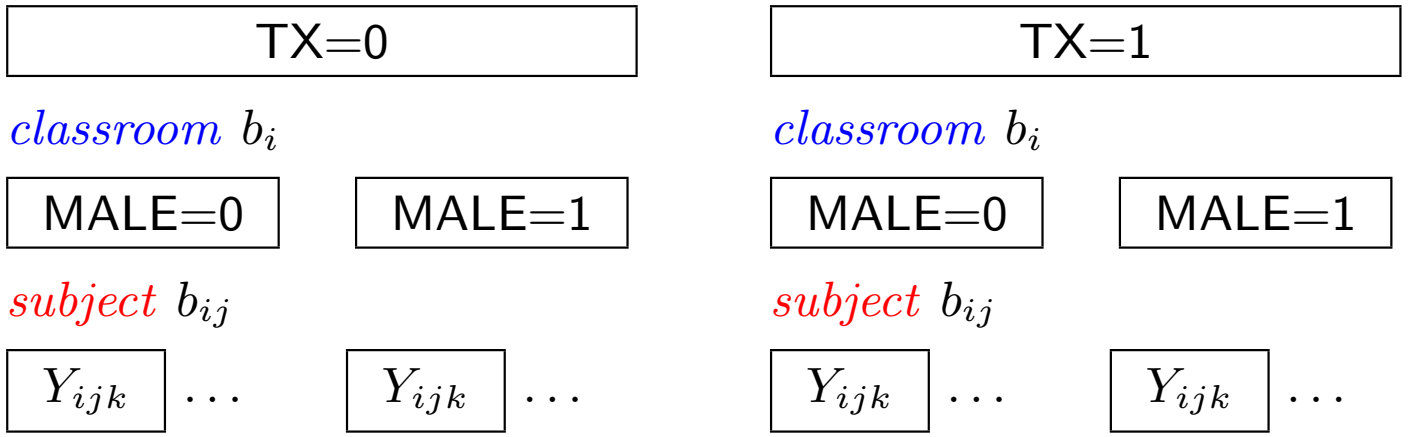
Multilevel Model Example

Scenario 3: (3 level)

Consider the following design:

- subjects clustered in classrooms
- repeated measurements per subject

Scenario 3



MARGINAL

$$\text{logit}E[Y_{ijk} \mid X_{ijk}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2ij}$$

CONDITIONAL

$$\text{logit}E[Y_{ijk} \mid X_{ijk},\ b_i,\ b_{ij}] = b_i + b_{ij} + \beta_0^{**} + \beta_1^{**} X_{1i} + \beta_2^{**} X_{2ij}$$

Comments:

- Marginal parameter, β, has simple interpretation that doesn't change as the "dependence" model changes.

- Conditional parameters, β^* and β^{**}, do not have the same interpretation as the marginal parameter.

- The interpretation of β^* is different than β^{**}. Is one of these "subject-specific"?

More Comments:

- The marginal parameter alone does not identify a complete probability model (except if observations are assumed independent).

- The conditional model does not allow identification of certain coefficients without assumptions on the distribution of b_{ij} (note connection to use of conditional logistic regression).

Table 3. Estimates for the multilevel model of modern prenatal care among women using some form of prenatal care†

Effects	*Results for the following methods:*							
	Logit	*MQL-1*	*MQL-2*	*PQL-1*	*PQL-2*	*PQL-B*	*Maximum likelihood*	*Gibbs*
Fixed effects								
Individual								
Child aged 3–4 years‡	−0.20	−0.17	−0.25	−0.22	−0.44	−0.81	−1.04	−1.33
Mother aged ⩾ 25 years‡	0.32	0.31	0.38	0.36	0.58	1.35	1.08	1.26
Birth order 2–3	−0.10	−0.10	−0.16	−0.13	−0.20	−0.49	−0.75	−1.00
Birth order 4–6	−0.23	−0.23	−0.32	−0.26	−0.31	−0.97	−0.56	−0.49
Birth order ⩾ 7	−0.19	−0.28	−0.45	−0.30	−0.45	−1.08	−1.08	−1.21
Family								
Indigenous, no Spanish‡	−0.84	−0.97	−1.02	−1.22	−2.18	−4.63	−5.60	−7.54
Indigenous Spanish‡	−0.57	−0.56	−0.93	−0.67	−1.00	−2.54	−2.62	−4.00
Mother's education primary‡	0.31	0.35	0.59	0.42	0.65	1.64	1.89	2.62
Mother's education secondary or better‡	1.01	0.90	1.06	0.98	1.93	3.81	3.61	5.68
Husband's education primary	0.18	0.22	0.32	0.25	0.30	0.95	0.96	1.11
Husband's education secondary or better‡	0.68	0.69	0.85	0.82	1.59	3.07	4.37	4.85
Husband's education missing	0.00	0.06	0.07	0.06	0.01	0.16	0.13	0.02
Husband professional, sales, clerk	−0.32	−0.40	−0.49	−0.47	−0.64	−0.60	−0.62	−0.56
Husband agricultural self-employed	−0.54	−0.52	−0.66	−0.62	−0.86	−1.75	−1.77	−2.64
Husband agricultural employee‡	−0.70	−0.27	−0.33	−0.29	−0.25	−2.34	−2.67	−3.77
Husband skilled service	−0.37	−0.15	−0.19	−0.18	−0.05	−1.05	−0.80	−1.12
Modern toilet in household‡	0.47	0.37	0.57	0.41	0.94	1.72	2.01	2.69
Television not watched daily	0.32	0.27	0.48	0.31	0.53	1.16	1.35	2.03
Television watched daily	0.47	0.33	0.41	0.39	0.67	1.55	1.51	2.05
Community								
Proportion indigenous, 1981‡	−0.90	−0.97	−1.61	−1.12	−2.05	−4.48	−5.01	−6.61
Distance to nearest clinic‡	−0.01	−0.01	−0.01	−0.01	−0.02	−0.05	−0.05	−0.07
Random effects								
Standard deviations σ								
Family	—	1.01	1.74	1.25	2.75	6.66	7.40	10.24
Community	—	0.79	1.23	0.86	1.71	3.48	3.74	5.40
Intraclass correlations ρ								
Family	—	0.33	0.58	0.41	0.76	0.95	0.95	0.98
Community	—	0.13	0.19	0.13	0.21	0.20	0.19	0.21

†The reference categories are child aged 0–2 years, mother's age less than 25 years, birth order 1, Ladino, mother no education, husband no education, husband not working or unskilled occupation, no modern toilet in the household and no television in the household.
‡Fixed effects are significant at the 5% level according to the maximum likelihood analysis.

TABLE 4. ESTIMATED ODDS RATIOS AND *t*-VALUES FOR MULTILEVEL LOGISTIC MODELS[a] OF THE PROBABILITY OF RECEIVING ANY PRENATAL CARE AND THE PROBABILITY OF RECEIVING FORMAL PRENATAL CARE AMONG THOSE WHO RECEIVED SOME CARE

	Any Prenatal Care (N = 3,409)		Formal Prenatal Care, Given Any (N = 2,449)	
Covariates	Odds Ratio[b]	*t*-value	Odds Ratio	*t*-value
Ethnicity				
(Ladino)				
Indigenous, no Spanish	1.53	0.51	0.004*	−3.21
Indigenous, Spanish	2.26	1.16	0.07*	−2.56
Individual Characteristics				
(Child age 0–2)				
Child age 3–4	0.57*	−3.08	0.35*	−3.25
(Mother < age 25)				
Mother age 25+	1.55	1.32	2.94*	2.02
(Birth order 1)				
Birth order 2–3	0.56	−1.89	0.47	−1.61
Birth order 4–6	0.33*	−2.60	0.57	−0.86
Birth order 7–16	0.19*	−3.12	0.34	−1.20
Socioeconomic Characteristics				
(Mother no education)				
Mother primary education	3.57*	3.28	6.63*	2.90
Mother secondary + education	33.21*	2.95	37.14*	2.52
(Husband no education)				
Husband primary education	1.77	1.42	2.60	1.50
Husband secondary + education	3.02	1.19	79.36*	2.40
Missing information	6.52*	2.92	1.14	0.13
(Husband no or unskilled occupation)				
Husband professional, sales, clerk	3.56	1.20	0.54	−0.41
Husband in agriculture, self-employed	0.80	−0.26	0.17	−1.39
Husband in agriculture, employed by others	1.02	0.02	0.07*	−2.07
Husband skilled, service	0.71	−0.38	0.45	−0.63
(No modern toilet in household)				
Modern toilet in household	2.90	1.66	7.43*	2.08
(No TV in household)				
TV, not watched daily	2.63	1.12	3.87	1.07
TV, watched daily	3.94*	2.18	4.54	1.61
Community Characteristics				
Proportion indigenous (1981)	1.80	0.57	0.007*	−2.99
Distance to nearest clinic (km)	0.98*	−2.35	0.95*	−3.33
σ_c	2.36*	8.00	3.74*	6.02
σ_l	4.84*	11.46	7.40*	6.10
ρ_c	0.22		0.26	
ρ_l	0.69		0.77	

*$p \leq .05$

Even More Comments:

- In the conditional models, no **direct** observation of the regression contrasts are available for covariates that vary slower than the random effects (ie. TX, and GENDER in the 3-level model).

- For descriptive contrasts it matters where the covariate IS relative to the random effects.

- For predictive contrasts it matters where the covariate COULD BE. (note difference between TX and GENDER).

- Holland (1986) discusses causal effects and claims that such contrasts are not appropriate for "attributes".

Estimation: β and/or b_i

Q: How to estimate β and/or b_i's?

- Jointly estimate β and b_i's (bias!)

- Parameterize b_i and then integrate over the distribution of the random effects (later)

- Eliminate b_i as nuisance parameters using a conditional likelihood

Consider simple paired data (Y_{i0}, Y_{i1}) with a "pre/post" covariate $\boldsymbol{X}_i = (X_{i0} = 0, X_{i1} = 1)$. Consider the logistic regression model:

$$\text{logit}(\mu_{ij}^b) \quad = \quad b_i + \beta_1 X_{ij}$$

Conditional Logistic Regression

	$Y_{i1}=1$	$Y_{i1}=0$
$Y_{i0}=1$	$\frac{\exp(b_i)}{[1+\exp(b_i)]} \cdot \frac{\exp(b_i+\beta)}{[1+\exp(b_i+\beta)]}$	$\frac{\exp(b_i)}{[1+\exp(b_i)]} \cdot \frac{1}{[1+\exp(b_i+\beta)]}$
$Y_{i0}=0$	$\frac{1}{[1+\exp(b_i)]} \cdot \frac{\exp(b_i+\beta)}{[1+\exp(b_i+\beta)]}$	$\frac{1}{[1+\exp(b_i)]} \cdot \frac{1}{[1+\exp(b_i+\beta)]}$

- We condition on the sum: $S_i = (Y_{i0} + Y_{i1})$, (known as a *sufficient statistic* for b_i)

- The sufficient statistic S_i only takes the values $0, 1, 2$.

- The conditional distribution of (Y_{i0}, Y_{i1}) is degenerate if $S_i = 0$ or $S_i = 2$.

- The only "informative" case is when $S_i = 1$.

$$P(Y_{i0}, Y_{i1} \mid S_i = 1) = \pi_{01}^{(1-Y_{i0})Y_{i1}} \pi_{10}^{Y_{i0}(1-Y_{i1})}$$

$$\begin{aligned} \pi_{01} &= P(Y_{i0} = 0, Y_{i1} = 1 \mid S_i = 1) \\ &= \frac{\exp(b_i + \beta)}{\exp(b_i) + \exp(b_i + \beta)} \\ &= \frac{\exp(\beta)}{1 + \exp(\beta)} \end{aligned}$$

$$\pi_{10} = \frac{1}{1 + \exp(\beta)}$$

- The conditional MLEs are:

$$\begin{aligned} \text{Let } A &= \sum_i \mathbf{1}(Y_{i0}=0, Y_{i1}=1) \\ \text{Let } B &= \sum_i \mathbf{1}(Y_{i0}=1, Y_{i1}=0) \end{aligned}$$

$$\begin{aligned} \widehat{\pi}_{01} &= A/(A+B) \\ \widehat{\beta} &= \log(A/B) \end{aligned}$$

- Connections to McNemar's test

- Connections to partial likelihood function

Conditional Likelihood and Cluster-level Covariates

- Suppose we extend the regression to include additional covariates:

$$\text{logit}(\mu_{ij}^b) = \underbrace{b_i + \boldsymbol{X}_{1i}\boldsymbol{\beta}_1}_{\text{between-cluster}} + \underbrace{X_{2ij}\boldsymbol{\beta}_2}_{\text{within-cluster}}$$

$$\begin{aligned}
\pi_{01} &= P(Y_{i0} = 0, Y_{i1} = 1 \mid S_i = 1) \\
&= \frac{\exp(b_i + \boldsymbol{X}_{1i}\boldsymbol{\beta}_1 + \boldsymbol{X}_{2i1}\boldsymbol{\beta}_2)}{\exp(b_i + \boldsymbol{X}_{1i}\boldsymbol{\beta}_1 + \boldsymbol{X}_{2i0}\boldsymbol{\beta}_2) + \exp(b_i + \boldsymbol{X}_{1i}\boldsymbol{\beta}_1 + \boldsymbol{X}_{2i1}\boldsymbol{\beta}_2)} \\
&= \frac{\exp[\boldsymbol{\beta}_2 \cdot (\boldsymbol{X}_{2i1} - \boldsymbol{X}_{2i0})]}{1 + \exp[\boldsymbol{\beta}_2 \cdot (\boldsymbol{X}_{2i1} - \boldsymbol{X}_{2i0})]}
\end{aligned}$$

Comments:

- The conditional likelihood eliminates $\boldsymbol{\beta}_1$ and b_i.

- For covariates that vary both between- and within- clusters the conditional likelihood only uses the information that comes from within-clusters.

- Extend to clusters with $n_i > 2$.

Example: VPS IC Analysis

```
*** [1] Baseline and Month 6 Only:  GEE ANALYSIS

xtgee q4safe ICgroup month6 ICgroupXmonth6 if month<=6, ///
  i(id) corr(exchangeable) family(binomial) link(logit) robust

*****************************************************************
*** Conditional Logistic Regression Analysis                  ***
*****************************************************************

*** [1] Baseline and Month 6 Only:  CONDITIONAL LOGISTIC

clogit q4safe ICgroup month6 ICgroupXmonth6 if month<=6, strata(id)
```

GEE Results

```
GEE population-averaged model                   Number of obs      =      2000
Group variable:                        id       Number of groups   =      1000
Link:                               logit       Obs per group: min =         2
Family:                          binomial                      avg =       2.0
Correlation:                exchangeable                      max =         2
                                                Wald chi2(3)       =     11.87
Scale parameter:                       1        Prob > chi2        =    0.0078
                          (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Semi-robust
      q4safe |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ICgroup |   .0162838   .1276727     0.13   0.899      -.23395    .2665177
      month6 |  -.0648189   .0985934    -0.66   0.511    -.2580585    .1284207
ICgroupXmo~6 |   .3664872   .1444608     2.54   0.011     .0833493    .6496251
       _cons |    .257412   .0902297     2.85   0.004     .0805651    .4342589
------------------------------------------------------------------------------
```

Conditional Logistic Regression Results

```
. clogit q4safe ICgroup month6 ICgroupXmonth6 if month<=6, strata(id)
note: multiple positive outcomes within groups encountered.
note: 692 groups (1384 obs) dropped due to all positive or
      all negative outcomes.
note: ICgroup omitted due to no within-group variance.

Conditional (fixed-effects) logistic regression   Number of obs   =        616
                                                  LR chi2(2)      =       8.60
                                                  Prob > chi2     =     0.0136
Log likelihood = -209.18813                       Pseudo R2       =     0.0201

------------------------------------------------------------------------------
      q4safe |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      month6 |  -.1082136   .1646397    -0.66   0.511    -.4309014    .2144743
ICgroupXmo~6 |   .5660467   .2311695     2.45   0.014     .1129628    1.019131
------------------------------------------------------------------------------
```

Comments

- The marginal coefficients are smaller in absolute value (ie. 0.366 versus 0.566 for `ICgroupXmonth6`).

- The marginal and conditional coefficients have different interpretations.

- The Z statistics, $\widehat{\beta}/s.e.$, are quite similar for the two regressions.

- Notice that in conditional logistic regression we can only estimate contrasts for within-cluster covariates **and** any interactions between a within-cluster covariate and a cluster-level covariate.

- See DHLZ sections 9.2 and 9.3 for additional detail.

Informed Consent: Waning?

GEE | Marginal mean

```
GEE population-averaged model                   Number of obs      =      3000
Group and time vars:                 id month   Number of groups   =      1000
Link:                                   logit   Obs per group: min =         3
Family:                               binomial                  avg =       3.0
Correlation:                     unstructured                   max =         3
                                (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
             |             Semi-robust
      q4safe |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ICgroup |   .0162838   .1276727     0.13   0.899      -.23395    .2665177
        post |  -.0648189   .0985934    -0.66   0.511    -.2580585    .1284207
     month12 |   .1466226   .1036148     1.42   0.157    -.0564587    .3497039
ICgroupXpost |   .3664872   .1444608     2.54   0.011     .0833493    .6496251
ICgroupXm~12 |  -.1204842   .1433102    -0.84   0.401     -.401367    .1603987
       _cons |    .257412   .0902297     2.85   0.004     .0805651    .4342589
------------------------------------------------------------------------------
```

Informed Consent: Waning?

Conditional Logistic Regression

```
. clogit q4safe ICgroup post month12 ICgroupXpost ICgroupXmonth12, strata(id)
note: multiple positive outcomes within groups encountered.
note: 524 groups (1572 obs) dropped due to all positive or
      all negative outcomes.
note: ICgroup omitted due to no within-group variance.

Conditional (fixed-effects) logistic regression   Number of obs    =       1428
                                                  LR chi2(4)       =      14.34
                                                  Prob > chi2      =     0.0063
Log likelihood = -515.76843                       Pseudo R2        =     0.0137
------------------------------------------------------------------------------
      q4safe |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |  -.0973385   .1560576    -0.62   0.533    -.4032058    .2085287
     month12 |   .2190657   .1563268     1.40   0.161    -.0873291    .5254606
ICgroupXpost |    .571471   .2262791     2.53   0.012     .1279722     1.01497
ICgroupXm~12 |  -.1786281   .2267166    -0.79   0.431    -.6229844    .2657283
------------------------------------------------------------------------------
```

Generalized Linear Mixed Models

Q: Are there alternatives to the use of conditional logistic regression that can explicitly parameterize heterogeneity yet estimate both β_1 and β_2?

A: Yes, Generalized Linear Mixed Models.

- Extend generalized linear models to correlated data!
- Extend linear mixed models to discrete outcome data!
- Likelihood estimation is computationally challenging
- "Mean" models are tangled with heterogeneity models

◦ Distributional assumptions?

◦ Scientific questions? Goals?

Review: Linear Mixed Models

Model

$$\boldsymbol{Y}_i = \underbrace{\boldsymbol{X}_i\boldsymbol{\beta}}_{\text{mean}} + \underbrace{\boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i}_{\text{covariance}}$$

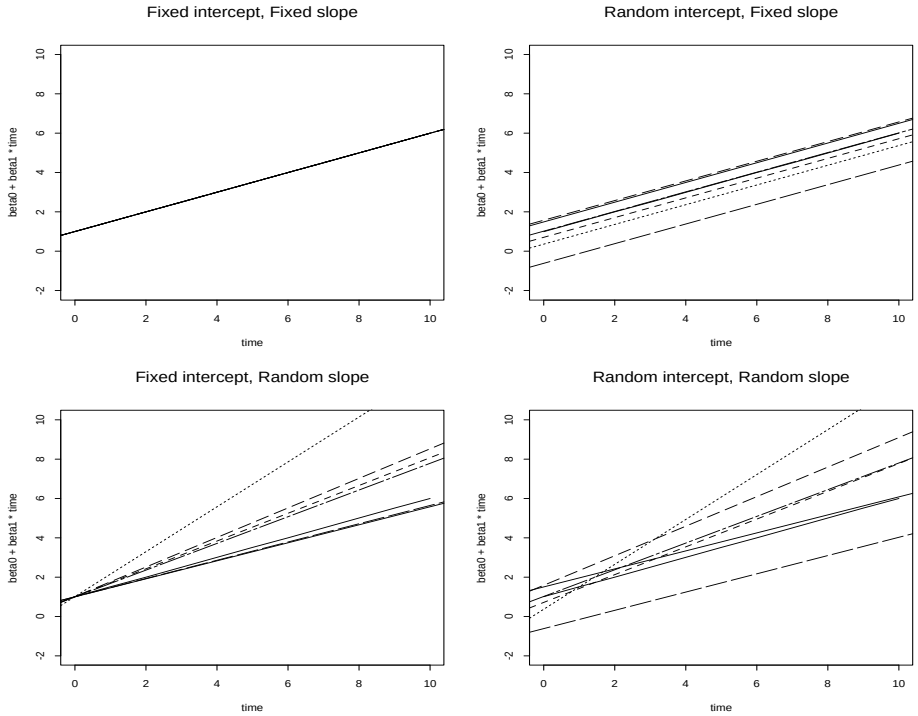
$$\begin{aligned} \boldsymbol{b}_i &\sim \mathcal{N}_q(\,\mathbf{0},\ \boldsymbol{D}[\boldsymbol{\alpha}]\,) \\ \boldsymbol{e}_i &\sim \mathcal{N}_{n_i}(\,\mathbf{0},\ \boldsymbol{R}_i[\boldsymbol{\alpha}]\,) \\ \boldsymbol{b}_i &\perp \boldsymbol{e}_i \end{aligned}$$

- $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$
- $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$

Estimation

- Maximum Likelihood (ML)
- REML

Fixed and Random Effects



Binary Data and Mixed Models

Q: Can we also use these mixed models for binary data?

A: Yes, but...

Model: Random Intercepts

$$P[Y_{ij} = 1 \mid \boldsymbol{X}_{ij}, b_i] = \pi_{ij}$$

$$\log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \boldsymbol{X}_{ij}\boldsymbol{\beta} + b_i$$

$$b_i \sim \mathcal{N}(0, \sigma_B^2)$$

Issues:

- Software

  NLMIXED (SAS)

  Stata (`xtlogit`, `gllamm`)

  BUGS

- Parameter interpretation issues

  Neuhaus, Kalbfleisch and Hauck (1991)

Generalized Linear Mixed Models

Model

- We build a hierarchical model, first specifying a GLM for Y_{ij} given the random effects:

$$\begin{aligned} \mu_{ij}^b &= E(Y_{ij} \mid \boldsymbol{X}_i, \boldsymbol{b}_i) \\ g(\mu_{ij}^b) &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i \\ [Y_{ij} \mid \boldsymbol{b}_i] &\sim \text{distribution} \\ Y_{ij} \perp Y_{ik} \mid \boldsymbol{b}_i &: \text{conditional independence} \end{aligned}$$

Generalized Linear Mixed Models

Model

- In the second stage (latent variable) we assume a population distribution for the "random effects"

$$\boldsymbol{b}_i \mid \boldsymbol{X}_i \quad \sim \quad \mathcal{N}_q(\, \mathbf{0},\ \boldsymbol{D}\,)$$

GLMMs: Estimation

- The likelihood function for the observed data, $\boldsymbol{Y}_i$, is obtained by integrating over the random effects distribution (latent variables, missing data).

- This integral is difficult to evaluate and has kept many statisticians busy finding ways to attack the integral!

- Modern computing power makes ML estimation feasible (although sometimes it can take a while).

- There are *approximate* ML methods (sometimes refered to as PQL or MQL), but we might not need these approximations since software is appearing.

$\star$

◦ **SAS** NLMIXED.

◦ **Stata** has logistic-normal routines.

◦ **BUGS** is a general purpose statistical software package that allows Bayesian inference.

Likelihood Evaluation

- Approximations:
  - Taylor series expansion around $b = 0$ (first order)
    - ⋆ Zeger, Liang & Albert (1988)
  - Laplace approximation: $E(b \mid \boldsymbol{Y})$
    - ⋆ Stiratelli, Laird & Ware (1984)
    - ⋆ Breslow and Clayton (1993)

Likelihood Evaluation

- Numerical Evaluation:
    - Gauss-Hermite quadrature
    - MCEM, MCNR
        - ⋆ McCulloch (1997)
        - ⋆ Booth and Hobert (1999)
        - ⋆ Hobert (2000)
- Bayes / MCMC:
    - Gibbs sampling
        - ⋆ Zeger and Karim (1991)

Example: Informed Consent

```
. xtlogit q4safe ICgroup post month12 ICgroupXpost ICgroupXmonth12, ///
   i(id) quad(20)

Random-effects logistic regression              Number of obs      =      3000
Group variable (i): id                          Number of groups   =      1000

Random effects u_i ~ Gaussian                   Obs per group: min =         3

                                                Wald chi2(5)       =     18.89
Log likelihood  = -1868.5603                    Prob > chi2        =    0.0020
------------------------------------------------------------------------------
      q4safe |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ICgroup |   .0249246    .195716     0.13   0.899    -.3586717    .4085209
        post |   -.099018   .1573788    -0.63   0.529    -.4074749    .2094388
     month12 |   .2237448   .1578929     1.42   0.156    -.0857196    .5332093
ICgroupXpost |   .5618163   .2255281     2.49   0.013     .1197893    1.003843
ICgroupXm~12 |  -.1839172   .2268745    -0.81   0.418    -.6285831    .2607487
       _cons |    .397937   .1385992     2.87   0.004     .1262876    .6695865
-------------+----------------------------------------------------------------
```

```
    /lnsig2u |   1.141153    .120744                      .9044989    1.377807
-------------+----------------------------------------------------------------
     sigma_u |   1.769287   .1068154                     1.571844     1.99153
         rho |   .4875789   .0301674                      .428898    .5466042
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   302.49 Prob >= chibar2 = 0.000
```

Summary: conditional models

• The GLM includes a term for the "cluster". This impacts our interpretation of the regression parameter $\boldsymbol{\beta}$.

◦ We may estimate the regression parameter using a conditional likelihood approach that *eliminates* the $\boldsymbol{b}_i$ by conditioning on their sufficient statistics.

◦ We may estimate the regression parameter using a marginal likelihood approach (ML) that *integrates* over the assumed distribution of $\boldsymbol{b}_i$ to obtain the marginal distribution of $\boldsymbol{Y}_i$.

◦ We may adopt a prior for the unknown parameters and proceed with a Bayesian analysis. MCMC and GS offer reasonable computational approaches to complex structure.

Example using SAS

```
options linesize=80 pagesize=60;

data hivnet;
  infile 'HivnetIC-SAS.data';
  input y visit ICgroup id month6 month12 post riskgp
        educ age cohort;
run;

proc genmod data=hivnet descending;
        class id riskgp;
model y = ICgroup post month12 ICgroup*post ICgroup*month12 /
                  dist=binomial link=logit;
        repeated subject=id / corrw type=un;
run;

proc nlmixed data=hivnet qpoints=20;
  parms B0 = 0.3000
        B_ICgroup = 0.0000
        B_post = 0.0000
        B_month12 = 0.0000
        B_IC_X_post = 0.0000
```

```
        B_IC_X_month12 = 0.0000
        sigma = 1.0;
  lp = B0 +
       B_ICgroup * ICgroup +
       B_post * post +
       B_month12 * month12 +
       B_IC_X_post * ICgroup*post +
       B_IC_X_month12 * ICgroup*month12;
  mu = exp( a + lp ) / (1 + exp( a + lp ) );
  model y ~ binomial( 1, mu );
  random a ~ normal( 0, sigma*sigma ) subject=id;
run;
```

GEE using GENMOD

The GENMOD Procedure

Model Information

Data Set	WORK.HIVNET
Distribution	Binomial
Link Function	Logit
Dependent Variable	y
Observations Used	3000

Response Profile

Ordered Value	y	Total Frequency
1	1	1775
2	0	1225

PROC GENMOD is modeling the probability that y='1'.

GEE Model Information

Correlation Structure	Unstructured
Subject Effect	id (1000 levels)
Number of Clusters	1000
Correlation Matrix Dimension	3
Maximum Cluster Size	3
Minimum Cluster Size	3

Working Correlation Matrix

	Col1	Col2	Col3
Row1	1.0000	0.3711	0.2751
Row2	0.3711	1.0000	0.3918
Row3	0.2751	0.3918	1.0000

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	0.2574	0.0902	0.0807	0.4342	2.85	0.0043
ICgroup	0.0163	0.1276	-0.2338	0.2664	0.13	0.8985
post	-0.0648	0.0985	-0.2580	0.1283	-0.66	0.5107
month12	0.1466	0.1036	-0.0564	0.3496	1.42	0.1568
ICgroup*post	0.3665	0.1444	0.0835	0.6495	2.54	0.0111
ICgroup*month12	-0.1205	0.1432	-0.4012	0.1603	-0.84	0.4003

GLMM Using NLMIXED

The NLMIXED Procedure

Specifications

Data Set	WORK.HIVNET
Dependent Variable	y
Distribution for Dependent Variable	Binomial
Random Effects	a
Distribution for Random Effects	Normal
Subject Variable	id
Optimization Technique	Dual Quasi-Newton
Integration Method	Adaptive Gaussian Quadrature

Dimensions

Observations Used	3000
Observations Not Used	0
Total Observations	3000
Subjects	1000
Max Obs Per Subject	3
Parameters	7
Quadrature Points	20

Parameters

B0	B_ICgroup	B_post	B_month12	B_IC_X_post	B_IC_X_month12	sigma
0.3	0	0	0	0	0	1

Parameters

NegLogLike

1916.68173

Iteration History

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	3	1892.22558	24.45615	78.81191	-1337
2	4	1870.62646	21.59912	11.45786	-578.234
3	6	1869.73699	0.88947	7.582643	-3.15352
4	7	1869.01284	0.724151	4.39984	-3.31256
5	9	1868.65382	0.359024	3.52137	-2.13052
6	11	1868.5722	0.08162	0.466648	-0.18294
7	13	1868.56783	0.004369	0.459775	-0.03191
8	14	1868.5626	0.005226	0.102357	-0.00944
9	16	1868.5625	0.000102	0.006276	-0.00017
10	18	1868.5625	6.567E-7	0.002089	-1.19E-6

NOTE: GCONV convergence criterion satisfied.

The NLMIXED Procedure

NOTE: GCONV convergence criterion satisfied.

Fit Statistics

-2 Log Likelihood	3737.1
AIC (smaller is better)	3751.1
AICC (smaller is better)	3751.2
BIC (smaller is better)	3785.5

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower
B0	0.3979	0.1386	999	2.87	0.0042	0.05	0.1259
B_ICgroup	0.02497	0.1957	999	0.13	0.8985	0.05	-0.3591
B_post	-0.09903	0.1574	999	-0.63	0.5293	0.05	-0.4079
B_month12	0.2237	0.1579	999	1.42	0.1568	0.05	-0.08609
B_IC_X_post	0.5618	0.2255	999	2.49	0.0129	0.05	0.1192
B_IC_X_month12	-0.1839	0.2269	999	-0.81	0.4177	0.05	-0.6291
sigma	1.7691	0.1067	999	16.58	<.0001	0.05	1.5596

Parameter Estimates

Parameter	Upper	Gradient
B0	0.6699	-0.002
B_ICgroup	0.4090	-0.00126
B_post	0.2098	-0.00209
B_month12	0.5336	-0.00101
B_IC_X_post	1.0044	-0.00111
B_IC_X_month12	0.2613	-0.00064
sigma	1.9785	-0.00009

Summary: NLMIXED

- The specification of the model requires writing the mean model explicitly using the logit (or other) function, and specification of the random effects.
- Maximum likelihood estimation is performed.
- Initial values for the parameters are required.
- **Q**: Interpretation of regression coefficients, β_j?
- **Q**: Interpretation of variance component, σ?

Response Conditional / Transition Models

- Another approach to summarizing / modelling multivariate categorical data is to model one response conditional on other responses. For example,

$$P(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \prod_{i=1} P(Y_{ij} \mid \mathcal{H}_{ij}, \boldsymbol{X}_i)$$

$$\mathcal{H}_{ij} = \mathcal{F}(Y_{i1}, Y_{i2}, \ldots, Y_{ij-1})$$

Example:

$$\text{logit}P(Y_6 = 1 \mid \texttt{ICgroup}, Y_0) = \beta_0 + \beta_1 \cdot \texttt{ICgroup} + \beta_2 \cdot Y_0 + \beta_3 \cdot \texttt{ICgroup} \cdot Y_0$$

$$\text{logit}P(Y_{12} = 1 \mid \texttt{ICgroup}, Y_6) = \beta_0 + \beta_1 \cdot \texttt{ICgroup} + \beta_2 \cdot Y_6 + \beta_3 \cdot \texttt{ICgroup} \cdot Y_6$$

- DHLZ Chapter 10 = "Transition Models"

```
***** Baseline to Month 6  *****

Call: glm(formula = q4safe6 ~ ICgroup * q4safe0,
                    family = binomial, data = vps.data)

Coefficients:
                     Value Std. Error    t value
    (Intercept) -0.7487170  0.1450553 -5.1615973
        ICgroup  0.5629999  0.1992980  2.8249147
        q4safe0  1.7101280  0.1968649  8.6868093
ICgroup:q4safe0 -0.2488770  0.2792883 -0.8911111

logOR for ICgroup=1 versus ICgroup=0 given Y0=0:  0.562 +/- 0.199

logOR for ICgroup=1 versus ICgroup=0 given Y0=1:  0.314 +/- 0.196
```

```
***** Month 6 to Month 12  *****

Call: glm(formula = q4safe12 ~ ICgroup * q4safe6,
                    family = binomial, data = vps.data)

Coefficients:
                    Value Std. Error    t value
    (Intercept) -0.4128452  0.1358824 -3.0382527
        ICgroup -0.1336985  0.2058819 -0.6493942
        q4safe6  1.4444462  0.1931056  7.4800849
ICgroup:q4safe6  0.5080077  0.2843519  1.7865459

logOR for ICgroup=1 versus ICgroup=0 given Y6=0:  -0.134 +/- 0.206

logOR for ICgroup=1 versus ICgroup=0 given Y6=1:   0.374 +/- 0.196
```

Transition Models

- These models nicely characterize $change$ for longitudinal categorical data.

- The likelihood factorization means that standard GLM software can be used to fit the models.

- The coefficients for the previous responses summarize the strength of longitudinal dependence (compare to our other models).

  ◦ Transition models do not directly model $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$.

Summary

- **Generalized Linear Mixed Models**
  - Model: GLM for $E(Y_{ij} \mid \boldsymbol{X}_i, \boldsymbol{b}_i)$
  - Conditional likelihood?
  - Observed data likelihood requires integration
    - + Approximations
    - + Quadrature
    - + MCMC
- **Transition Models**
  - Model: GLM for $E(Y_{ij} \mid \boldsymbol{X}_i, \mathcal{H}_{ij-1})$