Analysis of Longitudinal Data

- Patrick J. Heagerty PhD
- Department of Biostatistics
- University of Washington

Session Six Outline

- Generalized Linear Mixed Model (GLMM)
  - ▷ Specification of model
  - ▷ Interpretation of regression coefficients

- Estimation for GLMMs
  - ▷ Conditional likelihood
  - ▷ Maximum likelihood

GENERALIZED LINEAR MIXED MODELS (GLMMs)

Motivation

- Vaccine preparedness study (VPS), 1995-1998.
    - 5,000 subjects with high-risk for HIV acquisition.
    - Feasibility of phase III HIV vaccine trials.
    - Willingness, knowledge?

Motivation

- VPS Informed Consent Substudy (IC)
  - 20% selected to undergo mock informed consent.
  - Understanding of key items at 6mo, 12mo, 18mo.
- **Reference**: Coletti et al. (2003) *JAIDS*

Simple Example: VPS IC Analysis

To develop methods which assure that participants in future HIV vaccine trials understand the implications and potential risks of participating, the HIVNET developed a prototype informed consent process for a hypothetical future HIV vaccine efficacy trial. A 20% random subsample of the 4,892 Vaccine Preparedness Study (VPS) cohort was enrolled in a mock informed consent process at month 3 of the study (between the enrollment visit and the scheduled follow-up visit at month 6). Knowledge of 10 key HIV concepts and willingness to participate in future vaccine efficacy trials among these participants were compared with knowledge and willingness levels of participants not randomized to the informed consent procedure.

Simple Example: VPS IC Analysis

Items:

- Q4SAFE – “We can be sure that the HIV vaccine is safe once we begin phase III testing”

- NURSE – “The study nurse decides whether placebo or active product is given to a participant”

EDA – time cross-sectional

Baseline

```
ICgroup |q4safe0
        |0      |1      |RowTotl|
 -------+-------+-------+-------+
 0      |218    |282    |500    |
        |0.44   |0.56   |       |
 -------+-------+-------+-------+
 1      |216    |284    |500    |
        |0.43   |0.57   |       |
 -------+-------+-------+-------+
```

EDA – time cross-sectional

Post-Intervention, +3 months

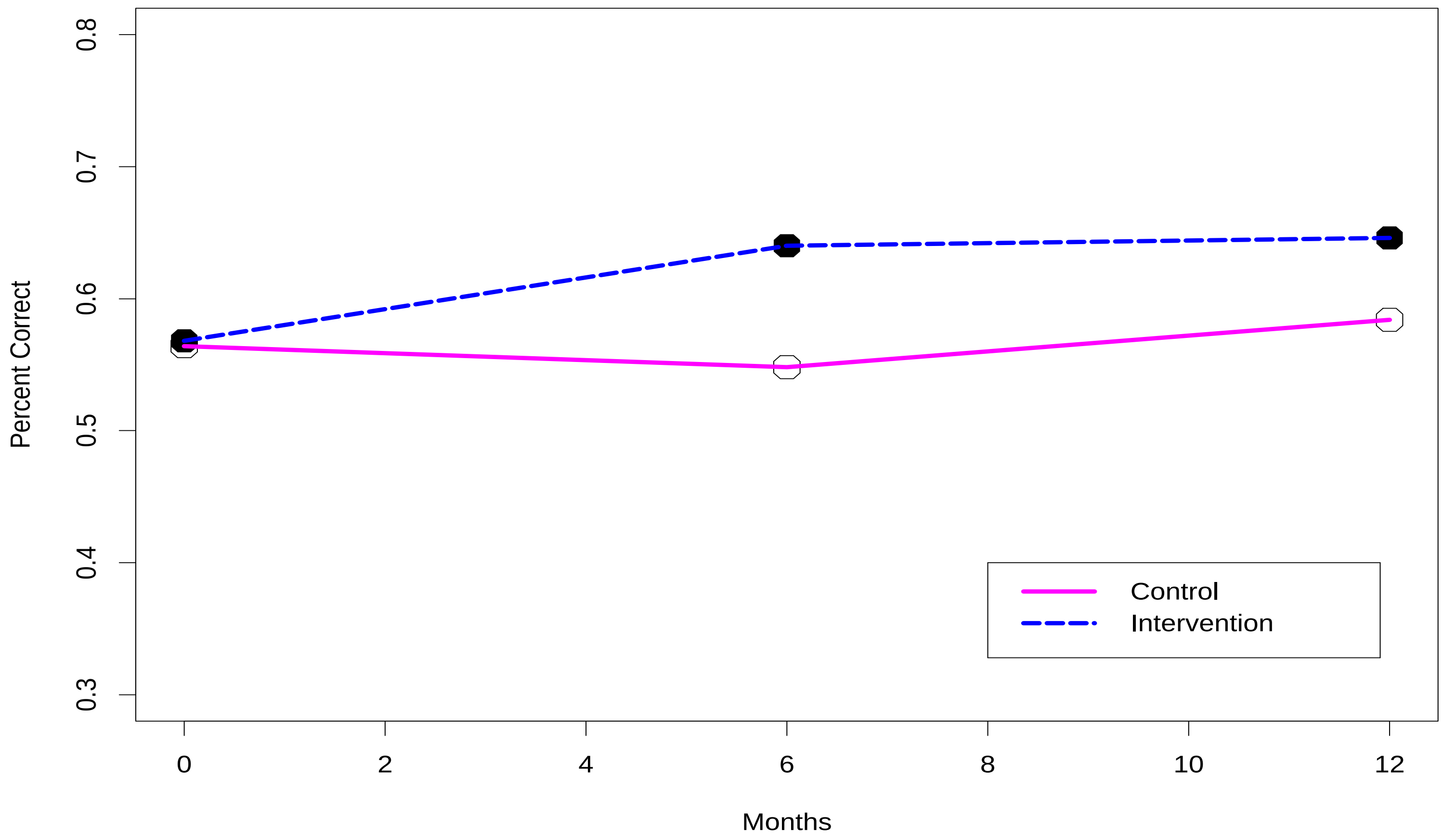
```
ICgroup |q4safe6
        |0       |1       |RowTotl|
 -------+--------+--------+-------+
 0      |226     |274     |500    |
        |0.45    |0.55    |       |
 -------+--------+--------+-------+
 1      |180     |320     |500    |
        |0.36    |0.64    |       |
 -------+--------+--------+-------+
```

EDA – time cross-sectional

Post-Intervention, +9 months

```
ICgroup |q4safe12
        |0      |1      |RowTotl|
 -------+-------+-------+-------+
 0      |208    |292    |500    |
        |0.42   |0.58   |       |
 -------+-------+-------+-------+
 1      |177    |323    |500    |
        |0.35   |0.65   |       |
 -------+-------+-------+-------+
```

HIVNET IC – Percent by Time and Group



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   |       |
-------+-------+-------+-------+
```

```
(a) Individual Change!

(b) Strong Association

      est OR = 5.53
```

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   |       |
-------+-------+-------+-------+
```

```
(a) Individual Change!

(b) Strong Association

      est OR = 4.30
```

Regression Models

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

Q: Does the intervention effect "wane"?

Regression Models:

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

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

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:

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

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

Conditional Regression Models

⋆⋆⋆ Assume that $[Y_{ij}, Y_{ik} \mid b_i] = [Y_{ij} \mid b_i][Y_{ik} \mid b_i]$ $\Rightarrow$ 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

- $\beta_2 \cdot$ (Time=6)

- $\beta_4 \cdot$ (Time=6 $\cdot$ Tx)

Parameter Interpretation

Between-cluster covariates

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

- 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)

Data: $Y_{ij} = 0/1$ infection status.

$X_{ij} = 0/1$ smoking status of mom.

* **cluster-level**: $X_{ij} \equiv X_i = 0/1$
* **observation-level**: $X_{ij} = 0/1$

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."

Example: Multilevel Logistic Model

- **Question(s)**: "Prenatal and Delivery Care ... in Guatemala: Do Family and Community Matter?" *Demography* (1996)
- **Data**:
  - **outcome(s)** = any prenatal care; formal care given any.
  - **covariates**:
    - **Child** characteristics (age, mom's age)
    - **Family** characteristics (ethnicity, education, occupation)
    - **Community** characteristics (percent indigenous, clinic distance)
- **Analysis**: logistic regression with Family and Community random effects.

Example: Multilevel Logistic Model

INDIGENOUS=0 | INDIGENOUS=1

community b_i | *community* b_i

LADINO=0 LADINO=1 | LADINO=0 LADINO=1

family b_{ij} | *family* b_{ij}

Y_{ijk} ... Y_{ijk} ... | Y_{ijk} ... Y_{ijk} ...

MARGINAL

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

CONDITIONAL

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

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

Covariates	Any Prenatal Care ($N = 3{,}409$)		Formal Prenatal Care, Given Any ($N = 2{,}449$)	
	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_f	4.84*	11.46	7.40*	6.10
ρ_c	0.22		0.26	
ρ_f	0.69		0.77	

*$p \leq .05$

Example: Multilevel Logistic Model

- These results suggest that GEE (with working independence) provides estimated odds ratios:
  - **Individual**: child (3-4yr) vs child (0-2yr)

    exp(-0.20) = 1/**1.22**

  - **Family**: indigenous vs ladino

    exp(-0.84) = 1/**2.32**

  - **Community**: indigenenous prop diff 10%

    exp[0.10*(-0.90)] = 1/**1.09**

Example: Multilevel Logistic Model

- These results suggest that a GLMM (with Family and Community random effects) provides estimated odds ratios:
  - ▷ **Individual**: child (3-4yr) vs child (0-2yr)

    exp(-1.04) = 1/**2.83**

  - ▷ **Family**: indigenous vs ladino

    exp(-5.60) = 1/**270**

  - ▷ **Community**: indigenenous prop diff 10%

    exp[0.10*(-5.01)] = 1/**1.65**

- Message: GEE and GLMM provide different estimates with different interpretations!

Estimation: β and/or b_i

Q: How to estimate $\boldsymbol{\beta}$ and/or b_i's?

- Jointly estimate $\boldsymbol{\beta}$ 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:

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

Review: Conditional Logistic Regression

- Conditional logistic regression is a method often introduced as appropriate for the analysis of "matched sets" of data.
  - ▷ e.g. paired subjects matched on AGE and HOSPITAL $(j = 0, 1)$

$$\text{logit}(\, p_{ij}\,) = \alpha(\text{AGE}_i, \text{HOSP}_i) + \beta \cdot X_{ij}$$

  - ▷ e.g. nested case-control data where match on TIME $(j = 0, 1, \ldots, 5)$

$$\text{logit}(\, p_{ij}\,) = \alpha(\text{TIME}_i) + \beta \cdot X_{ij}$$

Review: Conditional Logistic Regression

- Estimation of odds ratios while **controlling** for the matching factors (conditional likelihood).
- **No estimates** for the matching factors is provided.
- Estimation only uses “discordant pairs” (or sets).
- **Connections**:
  - ▷ We can view a matched set as a **cluster**.
  - ▷ We can view the matching criteria as corresponding to a cluster-effect, b_i (random or fixed).
    - ∗ e.g. $b_i = \alpha(\mathsf{AGE}_i, \mathsf{HOSP}_i)$
    - ∗ e.g. $b_i = \alpha(\mathsf{TIME}_i)$

Pre-Post Analysis using Conditional Logistic

- **Model:**
  - ▷ $Y_{ij} = 0/1$
  - ▷ $X_{ij} = 0/1$
    - ∗ Baseline (pre) time: $X_{ij} = 0$.
    - ∗ Follow-up (post) time: $X_{ij} = 1$.
  - ▷ logit(p_{ij}) $= b_i + \beta \cdot X_{ij}$

- **Estimation:** conditional likelihood that "eliminates" the b_i by conditioning on the sum of the Y_{ij}.
  - ▷ The next page considers the possible outcome probabilities

$$P[Y_{i0} = k_0, Y_{i1} = k_1 \mid X_{i0} = 0, X_{i1} = 1] = \pi_{k_0,k_1}$$

Conditional Logistic Regression

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

- 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:

$$\text{Let } A = \sum_i \mathbf{1}(Y_{i0} = 0, Y_{i1} = 1)$$

$$\text{Let } B = \sum_i \mathbf{1}(Y_{i0} = 1, Y_{i1} = 0)$$

$$\widehat{\pi}_{01} = A/(A+B)$$

$$\widehat{\beta} = \log(A/B)$$

- 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 $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$?

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?

Binary Data and Mixed Models

Model: Random Intercepts Logistic Regression

$$
\begin{aligned}
P[Y_{ij} = 1 \mid \boldsymbol{X}_{ij}, b_i] &= \pi_{ij} \\
\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) &= \boldsymbol{X}_{ij}\boldsymbol{\beta} + b_i \\
b_i &\sim \mathcal{N}(0, \sigma_B^2)
\end{aligned}
$$

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"

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

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.

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: GLMMs

• 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.