Chapter 13: Random effects models

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 770: Categorical Data Analysis

Chapter 13: Generalized Linear Mixed Models

Observations often occur in related clusters. Phrases like *repeated measures*, *longitudinal data*, and *panel data*, get at the same thing: there's correlation among observations in a cluster.

Chapter 12 dealt with an estimation procedure (GEE) that accounted for correlation in estimating population-averaged (marginal) effects.

This chapter models cluster correlation explicitly through *random effects*, yielding a GLMM.

Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT_i})$ be T_i correlated responses in cluster i. Associated with each repeated measure Y_{ij} are fixed (population) effects $\boldsymbol{\beta}$ and cluster-specific random effects $\mathbf{u}_i$. As usual, $\mu_{ij} = E(Y_{ij})$.

13.1.1 Add random effects to linear predictor

In a GLMM the linear predictor is augmented to include random effects:

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i.$$

for logistic regression, this is

$$\text{logit } P(Y_{ij} = 1) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i.$$

Note that conditional on $\mathbf{u}_i$,

$$E(Y_{ij}|\mathbf{u}_i) = \frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+\mathbf{z}'_{ij}\mathbf{u}_i}}{1 + e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+\mathbf{z}'_{ij}\mathbf{u}_i}}.$$

Example

I ask a random sample of *the same* $n = 22$ graduate students "do you like statistics?" once a month for 4 months.

$Y_{ij} = 1$ if "yes" and $Y_{ij} = 0$ if no. Here, $i = 1, \ldots, 30$ and $j = 1, \ldots, 4$.

Covariates might include m_{ij}, the average mood of the student over the previous month ($m_{ij} = 0$ is bad, $m_{ij} = 1$ is good), the degree being sought ($d_i = 0$ doctoral, $d_i = 1$ masters), the month $t_j = j$, and p_j the number of homework problems assigned in STAT 770 in the previous month.

A GLMM might be

$$\text{logit } P(Y_{ij} = 1) = \beta_0 + \beta_1 m_{ij} + \beta_2 d_i + \beta_3 p_j + \beta_4 j + u_i.$$

This model assumes that log-odds of liking statistics changes linearly in time, holding all else constant. Alternatively, we might fit a quadratic instead or treat time as categorical. Here, u_i represents a student's *a priori* disposition towards statistics.

Interpretation

Let's compare month $j+1$ to month j for individual i, holding all else (m, d, and p) constant. The difference in log odds is

$$(\beta_0 + \beta_1 m_{ij} + \beta_2 d_i + \beta_3 p_j + \beta_4 (j+1) + u_i) - (\beta_0 + \beta_1 m_{ij} + \beta_2 d_i + \beta_3 p_j + \beta_4 j + u_i) = \beta_4.$$

Not holding everything constant we get

$$(\beta_0 + \beta_1 m_{i,j+1} + \beta_2 d_i + \beta_3 p_{j+1} + \beta_4 (j+1) + u_i) - (\beta_0 + \beta_1 m_{ij} + \beta_2 d_i + \beta_3 p_j + \beta_4 j + u_i)$$

$$= \beta_1 (m_{i,j+1} - m_{ij}) + \beta_3 (p_{j+1} - p_j) + \beta_4.$$

Either way, we are conditioning on individual i, or *the subpopulation of all individuals with predisposition* u_i; i.e. everyone "like" individual i in terms of liking statistics to begin with.

How are e^{β_1}, e^{β_2}, e^{β_3} and e^{β_4} interpreted here?

Random effects $\mathbf{u}_1, \ldots, \mathbf{u}_n$

The random effects are assumed to come from (in general) a multivariate normal distribution

$$\mathbf{u}_1, \ldots, \mathbf{u}_n \overset{iid}{\sim} N_q(\mathbf{0}, \mathbf{\Sigma}).$$

The covariance $\text{cov}(\mathbf{u}_i) = \mathbf{\Sigma}$ can have special structure, e.g. exchangeability, AR(1), or be unstructured. The free elements of $\mathbf{\Sigma}$ are estimated along with $\boldsymbol{\beta}$.

- The $\mathbf{u}_i$ can account for heterogeneity caused by omitting explanatory variables.
- They can also explicitly model overdispersion, e.g.

$$Y_i \sim \text{Pois}(\lambda_i), \ \log \lambda_i = \mathbf{x}_i' \boldsymbol{\beta} + u_i, \ \ u_i \overset{iid}{\sim} N(0, \sigma^2).$$

Logit model for binary matched pairs

Recall $j = 1, 2$ denotes a binary covariate; for the PMA data it's time.

$$\text{logit } P(Y_{ij} = 1) = \alpha + u_i + \beta I\{j = 2\}.$$

Here, e^{β} is a cluster-specific odds ratio. We further assume $u_i \overset{iid}{\sim} N(0, \sigma^2)$.

Example: PMA data. Although a closed form estimate of β exists (see p. 494), we'll fit this in SAS using two different data structures for illustrative purposes.

```
data Data1;
 do ID=1 to 794; ap=1; time=0; output; ap=1; time=1; output; end;
 do ID=795 to 944; ap=1; time=0; output; ap=0; time=1; output; end;
 do ID=945 to 1030; ap=0; time=0; output; ap=1; time=1; output; end;
 do ID=1031 to 1600; ap=0; time=0; output; ap=0; time=1; output; end;
proc logistic data=Data1; strata ID;
 model ap(event='1')=time;
proc genmod data=Data1 descending; class ID;
 model ap=time / link=logit dist=bin;
 repeated subject=ID / corr=exch corrw;
```

Mixed model approach

On previous slide, first is conditional logistic approach from Chapter 11, second is marginal GEE logistic approach from Chapter 12.

Here is the GLMM approach of Chapter 13 with $u_i \overset{iid}{\sim} N(0, \sigma^2)$:

```
proc nlmixed maxiter=100 qpoints=100;
 parms beta0=1.0 beta1=-0.556 sigma=5.2;
 eta = beta0+beta1*time+u; pi = exp(eta)/(1+exp(eta));
 model ap ~ binary(pi);
 random u ~ normal(0,sigma*sigma) subject=ID;
 estimate 'subject-specific odds at 6 months' exp(beta1);
data matched;
input case occasion response count @@; datalines;
1  0 1  794     1  1 1  794      2  0 1  150      2  1 0  150
3  0 0   86     3  1 1   86      4  0 0  570      4  1 0  570
;
proc nlmixed maxiter=100 qpoints=100;
 eta = alpha + beta*occasion + u;  p = exp(eta)/(1 + exp(eta));
 model  response ~ binary(p);
 random u ~ normal(0, sigma*sigma) subject = case; replicate count;
```

SAS output

Output from the first fit:

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
beta0	1.2424	0.1857	1599	6.69	<.0001	0.05	0.8781	1.6067	-4.72E-7
beta1	-0.5563	0.1353	1599	-4.11	<.0001	0.05	-0.8216	-0.2910	-3.05E-7
sigma	5.1593	0.3527	1599	14.63	<.0001	0.05	4.4676	5.8510	8.779E-7

Additional Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower
subject-specific odds at 6 months	0.5733	0.07755	1599	7.39	<.0001	0.05	0.4212

Additional Estimates

Label	Upper
subject-specific odds at 6 months	0.7254

Read through **13.1.5**: random effects versus conditional approach.

13.2 Logistic normal model

A special, often-used case of the GLMM.

The logistic normal model is given by:

$$\text{logit } P(Y_{ij} = 1|u_i) = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i, \quad u_i \overset{iid}{\sim} N(0, \sigma^2).$$

When $\sigma = 0$ we get the standard logistic regression model, when $\sigma > 0$ we account for extra heterogeneity in clustered responses (each i is a cluster with it's own random u_i).

GLMMs induce only *positive* correlation within observations Y_{ij} and Y_{ik} within the same cluster.

13.2.3 Connection between marginal and conditional models

In the GEE approach, the marginal means are explicitly modeled:

$$\mu_{ij} = E(Y_{ij}) = g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta}),$$

and correlation among $(Y_{i1}, \ldots, Y_{iT_i})$ is accounted for in the estimation procedure.

The conditional approach models the means conditional on the random effects:

$$E(Y_{ij}|\mathbf{u}_i) = g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i).$$

The corresponding marginal mean is given by

$$E(Y_{ij}) = \int_{\mathbb{R}^q} g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i) f(\mathbf{u}_i; \boldsymbol{\Sigma}) d\mathbf{u}_i.$$

Marginal interpretation of logistic normal

In general, this is a complicated function of $\boldsymbol{\beta}$, however for the logistic-normal model when σ is "small," we obtain (not obvious)

$$E(Y_{ij}) \approx \exp(c\mathbf{x}'_{ij}\boldsymbol{\beta})/[1 + \exp(c\mathbf{x}'_{ij}\boldsymbol{\beta})],$$

where $c = 1/\sqrt{1 + 0.6\sigma^2}$. The *marginal odds* change by approximately $e^{c\beta_s}$ when x_{ijs} is increased by unity.

Since $c < 1$, the marginal effect is smaller than the conditional effect, reflecting that we are averaging with respect to the population. Note that the larger σ is, the more subject-to-subject variability there is, and the *smaller* the averaged effect $\hat{c}\hat{\beta}_s$ becomes.

Final look at PMA data

Here, $\hat{c} = 1/\sqrt{1 + 0.6(5.16)^2} = 0.24$. Then $e^{-0.556(0.24)} = 0.87$. Recall that the GEE approach yields $e^{-0.163} = 0.85$; not a bad approximation! Also recall that the conditional approach yielded $e^{-0.556}$. Severely annotated output:

The LOGISTIC Procedure, Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
time	1	-0.5563	0.1353	16.9152	<.0001

The GENMOD Procedure, Exchangeable Working Correlation

Correlation 0.7023650596

Analysis Of GEE Parameter Estimates, Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	0.3640	0.0508	0.2643	0.4636	7.16	<.0001
time	-0.1633	0.0390	-0.2398	-0.0868	-4.18	<.0001

The NLMIXED Procedure, Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
beta0	1.2424	0.1857	1599	6.69	<.0001	0.05	0.8781	1.6067	-4.72E-7
beta1	-0.5563	0.1353	1599	-4.11	<.0001	0.05	-0.8216	-0.2910	-3.05E-7
sigma	5.1593	0.3527	1599	14.63	<.0001	0.05	4.4676	5.8510	8.779E-7

13.2.4 Comments

- In epi studies, often want to compare disease prevalence across groups. Then it's of interest to compute marginal odds ratios and compare them.
- Direction and significance of effects usually the same across marginal/conditional models (e.g. PMA data).
- The more variability that's accounted for in the conditional model, the more we can "focus in" on the conditional effect of covariates. This is true in any situation where we block. This has the effect enlarging $\hat{\beta}_s$ estimates under a conditional model.
- When correlation is small, independence is approximately achieved, and marginal and conditional modeling yield similar results.
- GLMMs are being increasingly used, in part due to the availability of standard software to fit them!
- Bayesian approach is also natural here.

1. Opinion on legalized abortion (13.3.2)

	Response sequence							
Gender	(1,1,1)	(1,1,0)	(0,1,1)	(0,1,0)	(1,0,1)	(1,0,0)	(0,0,1)	(0,0,0)
Male	342	26	6	21	11	32	19	356
Female	440	25	14	18	14	47	22	547

Let (Y_{i1}, Y_{i2}, Y_{i3}) be the response to three questions asked of the same individual, "Do you support legalized abortion under three scenarios: (1) if the family has very low income, (2) the woman is unmarried & doesn't want to get married, (3) woman wants it for any reason?" $Y_{ij} = 1$ indicates "yes." A covariate of interest is gender: $x_i = 0$ for male $x_i = 1$ for female. A logistic-normal model is

$$\text{logit } P(Y_{ij} = 1) = \alpha + \beta_1 I\{j = 1\} + \beta_2 I\{j = 2\} + \gamma x_i + u_i, \quad u_i \overset{iid}{\sim} N(0, \sigma^2).$$

SAS data step

Within the same individual, e^{β_1} compares the odds of "yes" comparing "poor" to "any reason." e^{β_2} compares odds of "yes" comparing "single" to "any reason." $e^{\beta_2-\beta_1}$ compares odds of "yes" of "single" to "poor." e^{γ} compares odds of "yes" for females to males. Agresti's SAS code:

```
data new;
input sex poor single any count;
datalines;
1 1 1 1 342
1 1 1 0 26
1 1 0 1 11
1 1 0 0 32
1 0 1 1 6
1 0 1 0 21
1 0 0 1 19
1 0 0 0 356
2 1 1 1 440
2 1 1 0 25
2 1 0 1 14
2 1 0 0 47
2 0 1 1 14
2 0 1 0 18
2 0 0 1 22
2 0 0 0 457
;
```

SAS code

```
data new;  set new;
   sex = sex-1; case = _n_;
   q1=1; q2=0; resp = poor; output;
   q1=0; q2=1; resp = single; output;
   q1=0; q2=0; resp = any;  output;
drop poor single any;
proc nlmixed  qpoints = 50;
   parms alpha=0  beta1=.8  beta2=.3  gamma=0  sigma=8.6;
   eta = alpha + beta1*q1 + beta2*q2 + gamma*sex + u;
   p = exp(eta)/(1 + exp(eta));
   model resp ~ binary(p);
   random u ~ normal(0,sigma*sigma) subject = case;
   replicate count;
```

I added the following to get estimates of interest:

```
   estimate 'odds:   poor vs. any ' exp(beta1);
   estimate 'odds: single vs. any ' exp(beta2);
   estimate 'odds: single vs. poor' exp(beta2-beta1);
   estimate 'odds: female vs. male' exp(gamma);
```

Output

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
alpha	-0.6222	0.3812	1849	-1.63	0.1028	0.05	-1.3698	0.1255	0.000588
beta1	0.8358	0.1602	1849	5.22	<.0001	0.05	0.5217	1.1500	-0.0004
beta2	0.2929	0.1568	1849	1.87	0.0619	0.05	-0.01465	0.6004	0.000506
gamma	0.01272	0.4936	1849	0.03	0.9794	0.05	-0.9554	0.9809	0.000306
sigma	8.7878	0.5565	1849	15.79	<.0001	0.05	7.6964	9.8791	-0.00032

Additional Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
odds: poor vs. any	2.3068	0.3695	1849	6.24	<.0001	0.05	1.5821	3.0314
odds: single vs. any	1.3403	0.2102	1849	6.38	<.0001	0.05	0.9281	1.7525
odds: single vs. poor	0.5810	0.09137	1849	6.36	<.0001	0.05	0.4018	0.7602
odds: female vs. male	1.0128	0.5000	1849	2.03	0.0429	0.05	0.03226	1.9933

According to this (additive) model, there are significant differences within individuals on how they feel about legalized abortion depending on the circumstance. There is no significant difference due to gender. Under which circumstance is one's position on legalized abortion most favorable? Least?

Interpretation of GLMM and fit of marginal model

The estimate of $\hat{\sigma} = 8.8$ is quite large relative to the magnitude of the fixed effects (which are all less than unity). This reflects extreme heterogeneity in subject-to-subject response clusters (Y_{i1}, Y_{i2}, Y_{i3}). 1595 of 1850 subjects answered either $(0, 0, 0)$ or $(1, 1, 1)$. Does this also jibe with what we know about abortion as a "polarizing issue?"

Code to fit the marginal exchangeable model via GEE looks like:

```
data new; input sex poor single any count @@;
datalines;
1 1 1 1 342 1 1 1 0  26 1 1 0 1  11 1 1 0 0  32
1 0 1 1   6 1 0 1 0  21 1 0 0 1  19 1 0 0 0 356
2 1 1 1 440 2 1 1 0  25 2 1 0 1  14 2 1 0 0  47
2 0 1 1  14 2 0 1 0  18 2 0 0 1  22 2 0 0 0 457
;
data new;  set new;
 case=0; seq=_n_; * nesting case within sequence type (Y1,y2,y3);
 do i=1 to count;
  case=case+1;
  q1=1; q2=0; resp = poor;   output;
  q1=0; q2=1; resp = single; output;
  q1=0; q2=0; resp = any;    output;
 end;
 drop poor single any i count;
proc genmod; class case sex seq;
 model resp=q1 q2 sex / dist=bin link=logit;
 repeated subject=case(seq) / type=exch;
```

Nesting

This code makes use of nesting. Instead of having one case index $i = 1, \ldots, 1850$ for each individual, I have case nested within the type of sequence (Y_1, Y_2, Y_3), $i = 1, \ldots, j(i)$ where $j(1) = 342$, $j(2) = 26$, etc., $j(16) = 457$. This allows me to quickly get the data into a form SAS can use in PROC GENMOD. Output:

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	case(seq) (1850 levels)
Number of Clusters	1850
Correlation Matrix Dimension	3
Maximum Cluster Size	3
Minimum Cluster Size	3

Exchangeable Working Correlation

Correlation 0.8173308153

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-0.1219	0.0607	-0.2408	-0.0030	-2.01	0.0446
q1		0.1493	0.0297	0.0911	0.2076	5.02	<.0001
q2		0.0520	0.0270	-0.0010	0.1050	1.92	0.0544
sex	1	-0.0034	0.0878	-0.1756	0.1687	-0.04	0.9688
sex	2	0.0000	0.0000	0.0000	0.0000	.	.

Interpretation

As before, we see attenuation of the effects towards zero in the marginal model. From the conditional model we compute $\hat{c} = 1/\sqrt{1 + 0.6(8.79)^2} = 0.145$. Note that 0.15 is not too different from $0.12 = 0.145(0.836)$.

We can estimate the *population* ratio of odds for "poor" versus "single" by adding the command `estimate "odds poor vs. single" q1 1 q2 -1 / exp;` to the PROC GENMOD statement yielding:

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
odds poor vs. single	0.0973	0.0275	0.05	0.0434	0.1513	12.50	0.0004
Exp(odds poor vs. single)	1.1022	0.0303	0.05	1.0443	1.1633		

2. Longitudinal study of mental health (13.3.3)

Table 12.1 (p. 456) houses data from a longitudinal study comparing a new drug with a standard drug for treatment of subjects suffering mental depression. $n = 340$ Patients were either mildly or severely depressed upon admission into the study. At weeks 1, 2, and 4, corresponding to $j = 1, 2, 3$, patient i's suffering Y_{ij} was classified as normal $Y_{ij} = 1$ or abnormal $Y_{ij} = 0$. Let $s_i = 0, 1$ be the severity of the diagnosis (mild, severe) and $d_i = 0, 1$ denote the drug (standard, new).

We treat time as a categorical predictor and fit a marginal logit model with an exchangeable correlation structure:

```
data depress;
 infile "c:/tim/cat/depress.txt";
 input case diagnose treat time outcome; time=time+1;
proc genmod descending;  class case time;
  model outcome = diagnose treat time treat*time / dist=bin link=logit type3;
  repeated subject=case / type=exch corrw;
```

Output from marginal model

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	case (340 levels)
Number of Clusters	340
Correlation Matrix Dimension	3

Working Correlation Matrix

	Col1	Col2	Col3
Row1	1.0000	-0.0034	-0.0034
Row2	-0.0034	1.0000	-0.0034
Row3	-0.0034	-0.0034	1.0000

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		0.9812	0.1841	0.6203	1.3421	5.33	<.0001
diagnose		-1.3117	0.1453	-1.5964	-1.0269	-9.03	<.0001
treat		2.0427	0.3061	1.4428	2.6426	6.67	<.0001
time	1	-0.9601	0.2379	-1.4265	-0.4938	-4.04	<.0001
time	2	-0.6207	0.2372	-1.0855	-0.1559	-2.62	0.0089
time	3	0.0000	0.0000	0.0000	0.0000	.	.
treat*time	1	-2.0975	0.3923	-2.8663	-1.3287	-5.35	<.0001
treat*time	2	-1.0958	0.3900	-1.8602	-0.3314	-2.81	0.0050
treat*time	3	0.0000	0.0000	0.0000	0.0000	.	.

New is better than old

Score Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
diagnose	1	70.83	<.0001
treat	1	40.38	<.0001
time	2	15.73	0.0004
treat*time	2	29.52	<.0001

We see a severe diagnosis ($s = 1$) significantly decreases the odds of a normal classification by a factor of $e^{-1.31} = 0.27$.

The odds (or normal classification) ratio comparing the new drug to the standard drug changes with time because of the interaction. At 1 week it's $e^{2.04-2.09} = 0.95$, and week 2 it's $e^{2.04-1.10} = 2.6$, and at 4 weeks it's $e^{2.04-0} = 7.7$. The new drug is better, but takes time to work.

Next, a conditional, random effects model

Here, the focus is on whole populations of patients at 1, 2, and 4 weeks, and on the new drug versus the standard drug. These interpretations are not within the individual.

We now consider a conditional analysis

$$\begin{aligned} \text{logit } P(Y_{ij}=1) &= \alpha+\beta_1 s_i+\beta_2 d_i+\beta_3 I\{j=1\}+\beta_4 I\{j=2\} \\ &\quad +\beta_5 I\{j=1\}d_i+\beta_6 I\{j=2\}d_i+u_i \\ \text{where } u_i &\sim N(0,\sigma^2). \end{aligned}$$

I round parameter estimates from the GEE approach to use as starting values and fix `qpoints=200` (more on this later):

```
proc nlmixed  qpoints=200;
  parms a=1 b1=-1 b2=2 b3=-1 b4=-0.5 b5=-2 b6=-1 sig=.1;
  eta = a+b1*diag+b2*treat+b3*q1+b4*q2+b5*q1*treat+b6*q2*treat+u;
  p = exp(eta)/(1+exp(eta));
  model outcome ~ binary(p);
  random u ~ normal(0, sig*sig) subject=case;
```

NLMIXED output

The NLMIXED Procedure

AIC (smaller is better) 1176.8

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
a	0.9822	0.1844	339	5.33	<.0001	0.05	0.6194	1.3450	0.000363
b1	-1.3131	0.1543	339	-8.51	<.0001	0.05	-1.6165	-1.0097	0.000909
b2	2.0450	0.3129	339	6.54	<.0001	0.05	1.4296	2.6605	0.000101
b3	-0.9610	0.2313	339	-4.15	<.0001	0.05	-1.4160	-0.5060	-0.00049
b4	-0.6213	0.2256	339	-2.75	0.0062	0.05	-1.0650	-0.1775	0.000303
b5	-2.1002	0.3958	339	-5.31	<.0001	0.05	-2.8788	-1.3217	0.00004
b6	-1.0971	0.3852	339	-2.85	0.0047	0.05	-1.8548	-0.3394	-0.00046
sig	0.07027	1.1428	339	0.06	0.9510	0.05	-2.1777	2.3182	0.002123

The estimate $\hat{\sigma} = 0.07$ is small relative to the magnitude of the fixed effects. Let's refit the model without the random effects part:

```
proc nlmixed;
  parms a=1 b1=-1 b2=1 b3=-1.5 b4=-1 b5=-0.5 b6=-0.5;
  eta = a+b1*diag+b2*treat+b3*q1+b4*q2+b5*q1*diag+b6*q2*diag;
  p = exp(eta)/(1+exp(eta));
  model outcome ~ binary(p);
```

giving AIC

AIC (smaller is better) 1174.8

Independence model actually fits better

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
a	0.9812	0.1809	1020	5.43	<.0001	0.05	0.6263	1.3360	0.000029
b1	-1.3116	0.1462	1020	-8.97	<.0001	0.05	-1.5985	-1.0247	0.000048
b2	2.0430	0.3056	1020	6.68	<.0001	0.05	1.4432	2.6427	6.903E-6
b3	-0.9600	0.2290	1020	-4.19	<.0001	0.05	-1.4093	-0.5107	6.676E-6
b4	-0.6206	0.2245	1020	-2.76	0.0058	0.05	-1.0612	-0.1800	0.000017
b5	-2.0980	0.3893	1020	-5.39	<.0001	0.05	-2.8619	-1.3342	-4.79E-6
b6	-1.0961	0.3838	1020	-2.86	0.0044	0.05	-1.8491	-0.3431	0.000018

The AIC *drops* without the random effects! We have rather strong evidence that observations within a cluster (an individual here, taken at 1, 2, and 4 weeks) are essentially independent when adjusted for baseline covariates.

Note that the regression coefficients are essentially the same as those obtained from PROC GENMOD using the GEE approach. The absence of subject-to-subject heterogeneity implies that the marginal and conditional models are essentially the same.

13.6 Fitting binary GLMMs in PROC NLMIXED

The general model is hierarchical:

$$Y_{ij}|\mathbf{u}_i \overset{ind.}{\sim} \text{Bern}\left(\frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_i}}{1+e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_i}}\right),$$

$$\mathbf{u}_1,\ldots,\mathbf{u}_n \overset{iid}{\sim} N_q(\mathbf{0},\boldsymbol{\Sigma}).$$

Conditional on the random effect $\mathbf{u}_i$, the elements in $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{iT_i})$ are independent. So the PDF of $\mathbf{Y}_i|\mathbf{u}_i$ is

$$p(\mathbf{y}_i|\mathbf{u}_i) = \prod_{j=1}^{T_i}\left(\frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_i}}{1+e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_i}}\right)^{y_{ij}}\left(\frac{1}{1+e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_i}}\right)^{1-y_{ij}}.$$

However, the $\mathbf{u}_1,\ldots,\mathbf{u}_n$ are not model parameters. The model parameters are $(\boldsymbol{\beta},\boldsymbol{\Sigma})$. We need the to maximize the likelihood

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\Sigma}) = p(\mathbf{y}_1,\ldots,\mathbf{y}_n|\boldsymbol{\beta},\boldsymbol{\Sigma}).$$

Integrate to get likelihood

The *unconditional* PDF of $\mathbf{Y}_i$ is

$$p(\mathbf{y}_i) = \int_{\mathbb{R}^q} \left[\prod_{j=1}^{T_i} \frac{(e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_i})^{y_{ij}}}{1+e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_i}} \right] p(\mathbf{u}_i|\boldsymbol{\Sigma})\mathrm{d}\mathbf{u}_i,$$

where $p(\mathbf{u}_i|\boldsymbol{\Sigma})$ is a $N_q(\mathbf{0}, \boldsymbol{\Sigma})$ PDF. The $\mathbf{u}_i$ is integrated out and this is a function of $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ only. The likelihood is the product of these

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \int_{\mathbb{R}^q} \left[\prod_{j=1}^{T_i} \frac{(e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_i})^{y_{ij}}}{1+e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_i}} \right] p(\mathbf{u}_i|\boldsymbol{\Sigma})\mathrm{d}\mathbf{u}_i.$$

This involves n q-dimensional integrals that do not have closed-form.

Gauss-Hermite quadrature

PROC NLMIXED estimates the integrals (for a "current" quasi-Newton value of $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$) using adaptive Gauss-Hermite quadrature. This approach approximates the integrals above by sums

$$\int_{\mathbb{R}^q} h(\mathbf{u}_i) p(\mathbf{u}_i|\boldsymbol{\Sigma}) \mathrm{d}\mathbf{u}_i \approx \sum_{k=1}^{Q} c_k h(\mathbf{s}_k),$$

for arbitrary $h(\cdot)$ where Q is the number of quadrature points $\mathbf{s}_1, \ldots, \mathbf{s}_Q$ and $c_1, \ldots, c_Q$ are weights. The (adaptive) quadrature points and weights are chosen from a theory on integral approximations; we don't need to worry about that here.

Quasi-Newton maximization

Once the likelihood is approximated using quadrature, it is maximized via a quasi-Newton approach. The quasi-Newton approach does not require computing the matrix of second partial derivatives of the log-likelihood (the Hessian); rather this is approximated. Each iteration of the algorithm requires n integrals to be approximated! Suffice it to say, PROC NLMIXED can take awhile to run on large or complex data sets.

Note: there are other integral approximations SAS can use as well as other maximization procedures. I suggest reading the SAS documentation if you have trouble getting convergence of the algorithm for a particular model/data.

Tinkering with NLMIXED settings

There are two parameters to fool with when using "default" integral-approximation and maximization, `qpoints=`, the number of quadrature points, and `maxiter=`, the maximum number of quasi-Newton iterations to reach convergence criteria before you call the proceedings off.

Good starting values for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ can make or break the program, especially for large/complex data sets. You can try to guess starting values, or fit the model without random effects to get starting values for $\boldsymbol{\beta}$. I will often fit the Bayesian analogue to get starting values. Without a `parms=` statement in PROC NLMIXED, SAS gives all parameters ridiculous starting values of 1.

PMA data with default inputs

Optimization Technique	Dual Quasi-Newton
Integration Method	Adaptive Gaussian Quadrature
Quadrature Points	21

Iteration History

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	2	1991.81355	413.0301	198.8578	-6394.57
2	3	1855.73022	136.0833	101.5574	-203.759
3	5	1789.03422	66.696	28.79417	-68.5942
4	7	1782.45627	6.577942	7.531259	-4.88277
5	9	1781.67159	0.784685	1.447956	-0.68174
6	11	1781.59067	0.080917	0.640401	-0.08338
7	13	1781.58537	0.005299	0.204654	-0.00742
8	15	1781.58502	0.000357	0.005534	-0.00061
9	17	1781.58502	5.954E-7	0.00003	-1.18E-6

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
beta0	1.0173	0.1473	1599	6.91	<.0001	0.05	0.7284	1.3062	0.00003
beta1	-0.5038	0.1281	1599	-3.93	<.0001	0.05	-0.7551	-0.2524	0.000016
sigma	4.0151	0.2289	1599	17.54	<.0001	0.05	3.5662	4.4641	0.000012

Do defaults give good results?

Hmmm...“convergence criterion satisfied” seems to indicate everything's okay...or *is* it? Let's change the default code

```
proc nlmixed;
 eta = beta0+beta1*time+u; pi = exp(eta)/(1+exp(eta));
 model ap ~ binary(pi);
 random u ~ normal(0,sigma*sigma) subject=ID;
```

by inserting starting values based on the above estimates, i.e. adding:

```
 parms beta0=1.0 beta1=-0.5 sigma=4.0;
```

we obtain:

Iteration History

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	2	1753.63996	4.813275	8.496761	-78.1419
2	4	1752.03794	1.602023	6.378326	-66.1335
3	6	1751.66363	0.374302	6.784639	-7.27555
4	7	1751.03131	0.632326	1.980297	-1.24915
5	8	1750.91441	0.116899	0.190751	-0.21179
6	10	1750.91181	0.002596	0.044956	-0.00467
7	12	1750.91179	0.000021	0.002055	-0.00004
8	14	1750.91179	8.81E-8	0.000071	-1.65E-7

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
beta0	1.2540	0.1890	1599	6.63	<.0001	0.05	0.8832	1.6247	0.000071
beta1	-0.5576	0.1355	1599	-4.12	<.0001	0.05	-0.8233	-0.2919	4.6E-6
sigma	5.2073	0.3689	1599	14.12	<.0001	0.05	4.4837	5.9309	-0.00001

This is a bit different!

Both times we get the message "convergence criterion satisfied." What is happening? Answer: the likelihood is relatively flat around the MLE! So when we try PROC NLMIXED again with

```
parms beta0=1.25 beta1=-0.56 sigma=5.21;
```

the program crashes and in the log file we get:

```
ERROR: Quadrature accuracy of 0.000100 could not be achieved with 31 points.  The achieved
       accuracy was 0.000150.
```

We up the ante to `qpoints=100` and obtain:

Iteration History

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	4	1751.13016	0.004551	0.313641	-1.66969
2	7	1751.12844	0.001721	0.104651	-2.58751
3	10	1751.1281	0.000341	0.005987	-0.32286
4	11	1751.1281	5.023E-7	0.000027	-1.01E-6

NOTE: GCONV convergence criterion satisfied.

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
beta0	1.2424	0.1857	1599	6.69	<.0001	0.05	0.8781	1.6067	-0.00002
beta1	-0.5563	0.1353	1599	-4.11	<.0001	0.05	-0.8216	-0.2910	-0.00003
sigma	5.1593	0.3527	1599	14.63	<.0001	0.05	4.4676	5.8510	-0.00001

An approach that usually works

Now, if we had initially fit the model using the default `qpoints`, then put the resulting parameter estimates in as starting values but increase `qpoints=100`, we get the MLE immediately.

Parameters

beta0	beta1	sigma	NegLogLike
1	-0.5	4	1758.4624

Iteration History

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	2	1753.6971	4.765302	8.557791	-77.9394
2	4	1752.13848	1.558616	6.244724	-64.8866
3	6	1751.78562	0.352858	6.506276	-6.96529
4	7	1751.20889	0.576736	1.629794	-1.12045
5	8	1751.12946	0.079423	0.144148	-0.14538
6	10	1751.12811	0.001356	0.030861	-0.00248
7	12	1751.1281	8.376E-6	0.000783	-0.00002

NOTE: GCONV convergence criterion satisfied.

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
beta0	1.2424	0.1857	1599	6.69	<.0001	0.05	0.8781	1.6067	0.000461
beta1	-0.5563	0.1353	1599	-4.11	<.0001	0.05	-0.8216	-0.2910	0.000783
sigma	5.1593	0.3527	1599	14.63	<.0001	0.05	4.4675	5.8510	-0.00055

3. Clinical trial example (13.3.5)

Clinical trial with 8 centers; two creams compared to cure infection.

Center $Z = k$	Treatment X	Response Y		$\widehat{\theta}_{XY(k)}$
		Success	Failure	
1	Drug	11	25	1.2
	Control	10	27	
2	Drug	16	4	1.8
	Control	22	10	
3	Drug	14	5	4.8
	Control	7	12	
4	Drug	2	14	2.3
	Control	1	16	
5	Drug	6	11	∞
	Control	0	12	
6	Drug	1	10	∞
	Control	0	10	
7	Drug	1	4	2.0
	Control	1	8	
8	Drug	4	2	0.3
	Control	6	1	

Random effect u_i for each clinic

Center-to-center variability in how people respond to treatment can be incorporated in the conditional model

$$\text{logit } P(Y_{ij} = 1) = \alpha + \beta x_{ij} + u_i, \quad u_1, \ldots, u_8 \overset{iid}{\sim} N(0, \sigma^2),$$

where $x_{ij} = 0$ for drug and $x_{ij} = 1$ for control. SAS code:

```
data ctr1;
 input center$ treat s n @@; f=n-s; treat=treat-1;
 datalines;
 a 1 11 36 a 2 10 37  b 1 16 20 b 2 22 32
 c 1 14 19 c 2 7  19  d 1 2  16 d 2 1  17
 e 1 6  17 e 2 0  12  f 1 1  11 f 2 0  10
 g 1 1   5 g 2 1   9  h 1 4   6 h 2 6   7
;
data ctr2; set ctr1;
 do i=1 to n; if i<=s then y=1; else y=0; output; end;
proc nlmixed data=ctr2 qpoints=100;
 eta=alpha+beta*treat+u;
 p=exp(eta)/(1+exp(eta));
 model y ~ binary(p);
 random u ~ normal(0,sig*sig) subject=center;
```

SAS output & interpretation

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
alpha	-0.4591	0.5508	7	-0.83	0.4320	0.05	-1.7616	0.8433	0.000013
beta	-0.7385	0.3004	7	-2.46	0.0436	0.05	-1.4489	-0.02808	2.115E-6
sig	1.4008	0.4261	7	3.29	0.0133	0.05	0.3934	2.4083	0.000033

Within a given clinic, the odds of curing the infection is estimated to be (significantly) $1/e^{-0.739} = 2.1$ times greater on the drug versus the control. SAS will output empirical Bayes estimates of $u_1, \ldots, u_8$ by adding `out=re` (or whatever you want to call the new data set) to the `random` statement. Here they are:

Obs	center	Effect	Estimate	StdErr Pred	DF	tValue	Probt	Alpha	Lower	Upper
1	a	u	-0.09886	0.57554	7	-0.17177	0.86848	0.05	-1.45980	1.26208
2	b	u	1.85011	0.60147	7	3.07598	0.01792	0.05	0.42786	3.27235
3	c	u	0.99147	0.60198	7	1.64702	0.14355	0.05	-0.43199	2.41493
4	d	u	-1.29471	0.69606	7	-1.86006	0.10520	0.05	-2.94062	0.35121
5	e	u	-0.55775	0.64815	7	-0.86052	0.41800	0.05	-2.09038	0.97488
6	f	u	-1.60169	0.81836	7	-1.95719	0.09120	0.05	-3.53681	0.33343
7	g	u	-0.70444	0.76815	7	-0.91706	0.38961	0.05	-2.52081	1.11194
8	h	u	1.73721	0.74864	7	2.32047	0.05336	0.05	-0.03306	3.50747

Which clinic has the best overall success? Is it signficant?

4. Insomnia study (13.4.2)

Randomized, double-blind clinical trial comparing active hypnotic drug with placebo in insomnia patients. Response is time in minutes to fall asleep before going to bed. Each person has (Y_{i1}, Y_{i2}, x_i) where $Y_{i1} = 1, 2, 3, 4$ denotes time to fall asleep at baseline and $Y_{i2} = 1, 2, 3, 4$ is time to fall asleep after two weeks of treatment on one of $x_i = 0$ placebo or $x_i = 1$ hypnotic.

	Time to falling asleep				
		Follow-up			
Treatment	Initial	< 20	$20 - 30$	$30 - 60$	> 60
Active	< 20	7	4	1	0
	$20 - 30$	11	5	2	2
	$30 - 60$	13	23	3	1
	> 60	9	17	13	8
Placebo	< 20	7	4	2	1
	$20 - 30$	14	5	1	0
	$30 - 60$	6	9	18	2
	> 60	4	11	14	22

Proportional odds model random effects

This is repeated measures data on an individual with ordinal outcomes. A natural model to consider is an extension of the proportional odds model with a random effect that accounts for an individual's predisposition toward insomnia:

$$\text{logit } P(Y_{ij} \leq k|u_i) = \alpha_k + \beta_1 x_i + \beta_2 I\{j = 2\} + \beta_3 x_i I\{j = 2\} + u_i.$$

We are primarily interested in how the odds of taking less time to get to sleep changes from drug to placebo after being treated for two weeks (so $j = 2$). For $x_i = 1$,

$$\text{logit } P(Y_{i2} \leq k|u_i) = \alpha_k + \beta_1 + \beta_2 + \beta_3 + u_i,$$

for $x_i = 0$ we have

$$\text{logit } P(Y_{i2} \leq k|u_i) = \alpha_k + \beta_2 + u_i.$$

Building likelihood by hand

The difference of these is

$$\log\left\{\frac{P(Y_{i2}\leq k|x_i=1)/P(Y_{i2}>k|x_i=1)}{P(Y_{i2}\leq k|x_i=0)/P(Y_{i2}>k|x_i=0)}\right\}=\beta_1+\beta_3.$$

The likelihood, *conditional on the* u_i, is built from multinomial probabilities:

$$\begin{aligned}
P(Y_{ij}=1) &= P(Y_{ij}\leq 1)\\
P(Y_{ij}=2) &= P(Y_{ij}\leq 2)-P(Y_{ij}\leq 1)\\
P(Y_{ij}=3) &= P(Y_{ij}\leq 3)-P(Y_{ij}\leq 2)\\
P(Y_{ij}=4) &= 1-P(Y_{ij}\leq 3)
\end{aligned}$$

where

$$P(Y_{ij}\leq k)=\frac{e^{\alpha_k+\beta_1 x_i+\beta_2 I\{j=2\}+\beta_3 x_i I\{j=2\}+u_i}}{1+e^{\alpha_k+\beta_1 x_i+\beta_2 I\{j=2\}+\beta_3 x_i I\{j=2\}+u_i}}.$$

Code for fitting this model, adapted from Agresti's website

```
data insomnia;
 input case treat time outcome;
 y1=0; y2=0; y3=0; y4=0;
 if outcome=1 then y1=1;
 if outcome=2 then y2=1;
 if outcome=3 then y3=1;
 if outcome=4 then y4=1;
datalines;
        1       1     0        1
        1       1     1        1
        2       1     0        1
        2       1     1        1
etc...
      238       0     0        4
      238       0     1        4
      239       0     0        4
      239       0     1        4
;
proc nlmixed  qpoints=40;
  bounds i2 > 0;  bounds i3 > 0;
  eta1 = i1 + treat*beta1 + time*beta2 + treat*time*beta3 + u;
  eta2 = i1 + i2 + treat*beta1 + time*beta2 + treat*time*beta3 + u;
  eta3 = i1 + i2 + i3 + treat*beta1 + time*beta2 + treat*time*beta3 + u;
  p1 = exp(eta1)/(1 + exp(eta1));
  p2 = exp(eta2)/(1 + exp(eta2)) - exp(eta1)/(1 + exp(eta1));
  p3 = exp(eta3)/(1 + exp(eta3)) - exp(eta2)/(1 + exp(eta2));
  p4 = 1 - exp(eta3)/(1 + exp(eta3));
  ll = y1*log(p1) + y2*log(p2) + y3*log(p3) + y4*log(p4);
  model y1 ~ general(ll);
  estimate 'interc2' i1+i2; * this is alpha_2 in model, and i1 is alpha_1;
  estimate 'interc3' i1+i2+i3;  * this is alpha_3 in model;
  estimate 'd vs p at 2 weeks ' exp(beta1+beta3);
  estimate 'd vs p at baseline' exp(beta1);
  random u ~ normal(0, sigma*sigma) subject=case;
```

Output

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
i2	2.0050	0.1948	238	10.29	<.0001	0.05	1.6213	2.3886	0.000013
i3	2.0459	0.1942	238	10.54	<.0001	0.05	1.6634	2.4284	0.000012
i1	-3.4896	0.3588	238	-9.73	<.0001	0.05	-4.1964	-2.7828	0.000018
beta1	0.05786	0.3663	238	0.16	0.8746	0.05	-0.6637	0.7795	0.000022
beta2	1.6016	0.2834	238	5.65	<.0001	0.05	1.0434	2.1598	7.115E-7
beta3	1.0813	0.3805	238	2.84	0.0049	0.05	0.3318	1.8308	3.89E-6
sigma	1.9047	0.2314	238	8.23	<.0001	0.05	1.4489	2.3606	-7.43E-6

Additional Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
interc2	-1.4846	0.2903	238	-5.11	<.0001	0.05	-2.0566	-0.9127
interc3	0.5613	0.2702	238	2.08	0.0388	0.05	0.02909	1.0935
d vs p at 2 weeks	3.1241	1.1456	238	2.73	0.0069	0.05	0.8674	5.3808
d vs p at baseline	1.0596	0.3881	238	2.73	0.0068	0.05	0.2950	1.8241

Interpretation

The CI for $e^{\beta_1+\beta_3}$ is $(0.9, 5.4)$. We estimate the odds of falling asleep more quickly *after two weeks* is 3.1 times greater under the hypnotic for a randomly selected individual, but this is not statistically significant. At baseline the odds ratio $e^{\hat{\beta}_1}$ is 1.1.

We can also look at how the odds of falling to sleep 'earlier' changes from baseline to two weeks later by estimating e^{β_2} for placebo and $e^{\beta_2+\beta_3}$ for treatment:

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
2w vs base: placebo	4.9609	1.4057	238	3.53	0.0005	0.05	2.1916	7.7301
2w vs base: drug	14.6271	4.6261	238	3.16	0.0018	0.05	5.5137	23.7404

What is happening here? Do you believe in the 'placebo effect?'

This approach explicitly models an individual's predisposition toward falling asleep quickly through u_i.

Another approach

Another approach simply includes Y_{i1} as a baseline covariate and models Y_{i2} using the standard proportional odds model. This would give what one could expect under the treatment given an initial value Y_{i1}. The SAS code

```
data insomnia;
 input treat initial outcome count @@;
 datalines;
1 1 1  7 1 1 2  4 1 1 3  1 1 1 4  0
1 2 1 11 1 2 2  5 1 2 3  2 1 2 4  2
1 3 1 13 1 3 2 23 1 3 3  3 1 3 4  1
1 4 1  9 1 4 2 17 1 4 3 13 1 4 4  8
0 1 1  7 0 1 2  4 0 1 3  2 0 1 4  1
0 2 1 14 0 2 2  5 0 2 3  1 0 2 4  0
0 3 1  6 0 3 2  9 0 3 3 18 0 3 4  2
0 4 1  4 0 4 2 11 0 4 3 14 0 4 4 22
;
run;
proc logistic; class initial outcome / param=ref;
 model outcome = initial treat initial*treat;
 freq count;
 contrast 'sleep=1' treat 1 initial*treat 1 0 0 / estimate=exp;
 contrast 'sleep=2' treat 1 initial*treat 0 1 0 / estimate=exp;
 contrast 'sleep=3' treat 1 initial*treat 0 0 1 / estimate=exp;
 contrast 'sleep=4' treat 1 initial*treat 0 0 0 / estimate=exp;
```

SAS output

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
initial	3	49.9192	<.0001
treat	1	11.8416	0.0006
treat*initial	3	9.4082	0.0243

The odds of getting to sleep more quickly $P(Y_{i2} \leq k)/P(Y_{i2} > k)$ changes with both the treatment and the initial level of sleeplessness Y_{i1}. Let's compare the hypnotic to the placebo across the four levels of sleeplessness using the output from the four contrast statements:

Contrast Rows Estimation and Testing Results

Contrast	Type	Row	Estimate	Standard Error	Alpha	Confidence Limits		Wald Chi-Square	Pr > ChiSq
sleep=1	EXP	1	1.6963	1.2939	0.05	0.3804	7.5644	0.4800	0.4884
sleep=2	EXP	1	0.4295	0.2778	0.05	0.1209	1.5257	1.7076	0.1913
sleep=3	EXP	1	3.6747	1.5926	0.05	1.5716	8.5925	9.0185	0.0027
sleep=4	EXP	1	3.6910	1.4007	0.05	1.7544	7.7654	11.8416	0.0006

The odds of getting to sleep more quickly is significantly greater under the treatment for initial sleeplessness categories 3 and 4 (30-60 minutes and over 60 minutes).

GLMM example with bivariate random effects vectors

5. Poisson regression with multivariate random effects

Thall and Vail (1990) presented data from a clinical trial of $n = 59$ epileptic patients who were randomized to take either a new drug $d_i = 1$ or a placebo ($d_i = 0$) in addition to standard chemotherapy. Other baseline data included $a_i = \log(\text{age}_i)$ where age_i is age in years and $b_i = \log(\text{base}_i/4)$, where base_i is number of seizures in preceding 8-week period. The outcome is Y_{ij} the number of seizures within the following 4 2-week periods, up to 8 weeks. So $j = 1, 2, 3, 4$. The time variable used is actually $t_j = 0.2(j - 2.5)$. The model fit is

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij}),$$

where

$$\begin{aligned}
\log \lambda_{ij} &= \beta_0 + \beta_b b_i + \beta_d d_i + \beta_{bd} b_i d_i + \beta_a a_i + \beta_v t_j + u_{i1} + u_{i2} t_j \\
&= \begin{bmatrix} 1 \\ b_i \\ d_i \\ b_i d_i \\ a_i \\ t_j \end{bmatrix}' \begin{bmatrix} \beta_0 \\ \beta_b \\ \beta_d \\ \beta_{bd} \\ \beta_a \\ \beta_v \end{bmatrix} + \begin{bmatrix} 1 \\ t_j \end{bmatrix}' \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \\
&= \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i
\end{aligned}$$

Model building, continued...

We further assume

$$\mathbf{u}_1, \ldots, \mathbf{u}_{59} \overset{iid}{\sim} N_2(\mathbf{0}, \mathbf{\Sigma}) = N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}\right).$$

This assumes a log linear trend individual i's seizure rate over the 8 weeks. Specifically,

$$\log \lambda_{ij} = \theta_{0i} + \theta_{1i} \text{weeks}_j,$$

where $\text{cov}(\theta_{0i}, \theta_{1i}) = \sigma_{12}$. This follows from properties of multivariate normal distributions.

SAS code

```
data seiz1;
 input id$ seiz visit treat age base;
 age=log(age);
 base=log(base/4);
 if visit=1 then visit=-3;
 if visit=2 then visit=-1;
 if visit=3 then visit=1;
 if visit=4 then visit=3;
 visit=visit/10;
 datalines;
101   11    1   1   18   76
101   14    2   1   18   76
101   9     3   1   18   76
101   8     4   1   18   76
102   8     1   1   32   38
102   7     2   1   32   38
102   9     3   1   32   38
102   4     4   1   32   38
       ...et cetera...
238   13    1   0   22   47
238   15    2   0   22   47
238   13    3   0   22   47
238   12    4   0   22   47
;
proc nlmixed qpoints=50;
parms b_const=-1.3 b_base=0.9 b_trt=-0.9 b_basetrt=0.3 b_age=0.2
      b_visit=-0.3 s11=0.25 s22=0.53 s12=0.003;
eta=b_const+b_base*base+b_trt*treat+b_age*age+b_basetrt*base*treat
    +b_age*age+b_visit*visit+u1+u2*visit;
lambda=exp(eta);
model seiz ~ poisson(lambda);
random u1 u2 ~ normal([0,0],[s11,s12,s22]) subject=id;
```

Annotated output

Parameter	Estimate	SE	Pr > |t|	Lower	Upper
b_const	-1.3682	1.2007	0.2593	-3.7726	1.0363
b_base	0.8850	0.1313	<.0001	0.6221	1.1478
b_trt	-0.9287	0.4022	0.0246	-1.7340	-0.1233
b_basetrt	0.3380	0.2044	0.1038	-0.07142	0.7474
b_age	0.2384	0.1768	0.1830	-0.1157	0.5924
b_visit	-0.2664	0.1647	0.1113	-0.5962	0.06342
s11	0.2515	0.05879	<.0001	0.1338	0.3692
s22	0.5315	0.2294	0.0241	0.07214	0.9908
s12	0.002871	0.08870	0.9743	-0.1748	0.1805

Interpretation

Consider an individual from the population with covariates (a, b) at time t with random effect (u_1, u_2). The ratio of seizure rates, within this individual, for drug versus placebo is:

$$\frac{\lambda(a, b, t, d = 1|\mathbf{u})}{\lambda(a, b, t, d = 0|\mathbf{u})} = \frac{e^{\beta_0+\beta_b b+\beta_d+\beta_{bd} b+\beta_a a+\beta_v t+u_1+u_2 t}}{e^{\beta_0+\beta_b b+\beta_a a+\beta_v t+u_1+u_2 t}} = e^{\beta_d+\beta_{bd} b}.$$

Within a subject, the mean number of seizures over a 2-week period is reduced by $e^{-0.929+0.338 \log(\text{base}/4)} = (0.247)\text{base}^{0.338}$.

This function crosses unity between 62 and 63 baseline seizures within the previous 8 weeks. It's about 0.5 when base $= 8$. So the drug significantly reduces seizures at any visit, but the reduction rate critically depends on the baseline seizure rate.

Would any other interactions be of interest here? How about a visit by treatment interaction?

Some final comments

- 13.5 discusses multilevel modeling: different sets of random effects at different levels of a hierarchy (e.g. a student takes a battery of tests at a school: students within school, schools within state, state within country).
- Agresti has G^2 for GOF (deviance-based) tests when looking at contingency tables (e.g. leading crowd example). We get the maximized log-likelihood out of PROC NLMIXED. If careful we might be able to get G^2.
- Did not discuss Bayesian approaches; very natural here.
- Can check normality assumption by looking at $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_n$ but problems with this when cluster sizes are small.

More final comments

- 13.6.5 discusses testing $H_0 : \sigma = 0$ versus $H_1 : \sigma > 0$ in a simple model with univariate $u_1, \ldots, u_n \overset{iid}{\sim} N(0, \sigma^2)$. Fit the full model with random effects compute L_f (maximized log-likelihood), fit simpler model without random effects $\sigma = 0$ and get L_r. Let $t = -2[L_r - L_f]$ be the LRT statistic. The p-value for the test is $p = 0.5P(\chi^2_1 > t)$.
- 13.1.5: random intercept model not appropriate in case-control study as clusters not randomly sampled. Apparently there are fixes to this.
- Note that can have `model success~binomial(trials,prob);` in NLMIXED as well as other distributions; see the documentation. Useful homework problem 13.2.

PROC GLIMMIX

This is a recent SAS procedure that fits GLMM's.

- The procedure had been available as a macro for some time.
- GLIMMIX essentially extends the MIXED procedure to GLM's, and in fact iteratively calls MIXED when fitting GLMM's.
- Only normal random effects are allowed.
- GLIMMIX uses an approximation when fitting models. The approximation in effect replaces the intractable integral that NLMIXED approximates (using quadrature) with a simple linear Taylor's expansion. It's crude, but can work and it's fast. See SAS' GLIMMIX documentation for details on "Pseudo-likelihood Estimation Based on Linearization." Also described in your book in 13.6.4.

Model notation

- The model is $E\{\mathbf{Y}|\gamma\} = g^{-1}\{\mathbf{X}\beta + \mathbf{Z}\gamma\}$ where $\gamma \sim N_q(\mathbf{0}, \mathbf{\Sigma})$. Also, $var\{\mathbf{Y}|\gamma\} = \mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}$ where $\mathbf{A}^{1/2}$ comes from the sampling model (e.g. Poisson, normal, binomial) and $\mathbf{R}$ is a 'marginal' covariance matrix. For GLMM's, $\mathbf{R} = \phi\mathbf{I}$ and so $var\{\mathbf{Y}|\gamma\} = \phi\mathbf{A}$ where $\mathbf{A}$ is a diagonal matrix. This is because for GLMM's, the responses within a cluster are independent given the random effects γ.
- GLIMMIX can also fit marginal models allowing for correlation within a cluster (like GENMOD), but uses a different estimation method than GENMOD with the repeated statement. Then $\mathbf{R}$ has structure, e.g. exchangeability (called compound symmetry here), AR structure, spatial structures, and others found in PROC MIXED.
- The learning curve is steep, although it's nice to be aware of alternative fitting procedures if necessary!

Ache monkey hunting

Data on the number of capuchin monkeys killed by 47 Ache hunters over several hunting trips were recorded. There were 363 total records. I'll describe the hunting process in class; it involves splitting into groups, chasing monkeys through the trees, and shooting arrows straight up.

Let Y_{ij} be the number of monkey's killed by hunter i, $i = 1, \ldots, 47$ on trip j of length L_{ij} (the trip length serves as an 'offset' in the model fitting). Let λ_i be the hunter i's kill rate (per day).

$$Y_{ij} \sim \text{Poisson}(\lambda_i L_{ij}),$$

where

$$\log \lambda_i = \beta_0 + \beta_1 a_i + \beta_2 a_i^2 + u_i,$$

$$u_1, \ldots, u_{47} \overset{iid}{\sim} N(0, \sigma^2).$$

Monkey hunting, continued...

- Monkey hunting is dangerous.
- We include a quadratic effect because we expect a "leveling off" effect or possible decline in ability with age.
- Of interest is when hunting ability is greatest. Hunting prowess contributes to a man's status within the group. a_i is hunter i's age-45 years.
- An individual's kill rate is given by $\lambda = e^{\beta_0+\beta_1 a+\beta_2 a^2} e^u$, where a is the individual's age and u is their latent hunting ability.
- One can compare the effect of age within the span of, say, 20 to 60 years, to the spread of e^u to see which explains more of the variability in terms of hunting ability: age or innate ability.

Monkey hunting data step

Data sorted by trip number:

```
data ache1;
 input TRIP$ PID$ AGE nkills tripdays; ltripday=log(tripdays); age=age-45;
 datalines;
C082697A    3394    31  1   4
C082697A    3327    38  0   4
C082697A    3313    39  0   4
C082697A    3220    50  0   4
C082697A    3157    56  0   4
C082697A    3146    57  0   4
C082697A    3144    58  1   4
C082697A    7089    59  1   4
C082697A    3126    60  2   4
C082697A    7085    62  1   4
C102197A    3394    31  1   3
C102197A    3327    38  0   3
C102197A    3238    48  3   3
C102197A    3220    50  0   3
C102197A    3144    58  2   3
C102197A    3086    67  0   3

        ...et cetera...

T120997A    3182    53  0   5
T120997A    3094    65  0   5
T121597A    3254    46  0   4
T121597A    3128    60  0   4
;
```

Monkey hunting SAS code

With calls to genmod, nlmixed, and glimmix:

```
proc genmod data=ache1;
 class pid;
 model nkills=age age*age / dist=poisson link=log offset=ltripday;
 repeated subject=pid / type=exch;

proc glimmix data=ache1; class pid;
 model nkills = age age*age / dist=pois link=log offset=ltripday solution;
 random _residual_ / subject=pid type=cs;

proc sort; by pid; run; * need to sort by subject!;
proc nlmixed qpoints=100 data=ache1;
 parms b1=-2.3 b2=0.0251 b3=-0.002 v=1.0;
 eta=b1+b2*age+b3*age**2+u+ltripday;
 lambda=exp(eta);
 model nkills ~ poisson(lambda);
 random u ~ normal(0,v) subject=pid;

proc glimmix data=ache1; class pid;
 model nkills = age age*age / dist=pois link=log offset=ltripday solution;
 random intercept / subject=pid;
```

GENMOD GEE output

The GENMOD Procedure

Class	Levels	Values
PID	47	3086 3094 3111 3126 3128 3139 3144 3146 3157 3166 3172 3182 3217 3220 3238 3240 3254 3302 3313 3316 3322 3327 3349 3371 3378 3386 3390 3394 3401 3405 3414 3416 3434 3436 3450 3465 3480 3486 3495 3525 3529 3548 3572 7032 7085 7089 8024

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	PID (47 levels)
Number of Clusters	47
Correlation Matrix Dimension	28
Maximum Cluster Size	28
Minimum Cluster Size	1

Exchangeable Working Correlation

Correlation 0.2180742191

Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-2.2901	0.3266	-2.9303	-1.6500	-7.01	<.0001
AGE	0.0141	0.0257	-0.0363	0.0645	0.55	0.5840
AGE*AGE	-0.0019	0.0015	-0.0049	0.0010	-1.30	0.1937

GLIMMIX marginal output

Compare to GENMOD:

The GLIMMIX Procedure

Model Information

Response Variable	nkills
Response Distribution	Poisson
Link Function	Log
Variance Function	Default
Offset Variable	ltripday
Variance Matrix Blocked By	PID

Dimensions

R-side Cov. Parameters	2
Subjects (Blocks in V)	47
Max Obs per Subject	28

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
CS	PID	0.2730	0.1077
Residual		1.8348	0.1422

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-2.3253	0.2836	46	-8.20	<.0001
AGE	0.01589	0.01657	314	0.96	0.3385
AGE*AGE	-0.00181	0.001374	314	-1.31	0.1898

NLMIXED GLMM output

The NLMIXED Procedure

Specifications

Dependent Variable	nkills
Distribution for Dependent Variable	Poisson
Random Effects	u
Distribution for Random Effects	Normal
Subject Variable	PID
Optimization Technique	Dual Quasi-Newton
Integration Method	Adaptive Gaussian Quadrature

Dimensions

Total Observations	363
Subjects	47
Max Obs Per Subject	28
Parameters	4
Quadrature Points	100

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
b1	-2.6229	0.4515	46	-5.81	<.0001	0.05	-3.5317	-1.7141
b2	0.03385	0.02521	46	1.34	0.1859	0.05	-0.01689	0.08458
b3	-0.00491	0.002280	46	-2.16	0.0364	0.05	-0.00950	-0.00033
v	2.1081	0.8926	46	2.36	0.0225	0.05	0.3115	3.9048

GLIMMIX GLMM output

The GLIMMIX Procedure

Model Information

Response Variable	nkills
Response Distribution	Poisson
Link Function	Log
Variance Function	Default
Offset Variable	ltripday
Variance Matrix Blocked By	PID

Dimensions

G-side Cov. Parameters	1
Subjects (Blocks in V)	47
Max Obs per Subject	28

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	PID	1.7965	0.6505

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-2.4222	0.4113	46	-5.89	<.0001
AGE	0.02889	0.02307	314	1.25	0.2115
AGE*AGE	-0.00405	0.002079	314	-1.95	0.0525

Comments

- Notice the similarities in the GENMOD and GLIMMIX output fitting (the first two) marginal models.
- Notice the similarities in the GENMOD and GLIMMIX output fitting (the last two) conditional GLMM models.
- The quadratic effect is significant in the random effects models, but not the marginal models. This often happens when you focus on the individual.
- One benefit of fitting conditional random effects models: prediction is possible!

General comments

- Why GLIMMIX? To easily handle crossed or nested random effects. To handle large dimensional random effects. To jointly model counts and continuous outcomes. To avoid waiting 3 hours for NLMIXED to converge. To fit spatial covariance and other complex covariance structures with GLM's that cannot be accommodated by GENMOD.
- Why not GLIMMIX? It uses approximations which can bias results. You don't know how biased your results actually are. However, most models are approximations to reality to begin with so maybe not that big of a deal.
- Bayesian approach also natural but not as fast or easy to implement. However, no approximations are used and inference is exact up to Monte Carlo error.
- There are other packages out there to perform similar analyses.