

## 12.6 Fitting binary GLMMs in PROC NLMIXED

The general model is hierarchical:

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

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

Conditional on the random effect  $\mathbf{u}_i$ , the elements in  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT_i})$  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  $\mathbf{u}_1, \dots, \mathbf{u}_n$  are not model parameters. The model parameters are  $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ . We need to maximize the likelihood

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

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

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

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

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

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

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

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

for arbitrary  $h(\cdot)$  where  $Q$  is the number of quadrature points  $\mathbf{s}_1, \dots, \mathbf{s}_Q$  and  $c_1, \dots, c_Q$  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.

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.

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  $\beta$  and  $\Sigma$  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  $\beta$ . 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.



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			Pr >  t	Alpha	Lower	Upper	Gradient
		Error	DF	t Value					
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			Pr >  t	Alpha	Lower	Upper	Gradient
		Error	DF	t Value					
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

Now, if we had initially fit the model using the default `qpnts`, then put the resulting parameter estimates in as starting values but increase `qpnts=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** Clinical trial with 8 centers; two creams compared to cure infection.

Center $Z = k$	Treatment $X$	Response $Y$		$\hat{\theta}_{XY(k)}$
		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	$\infty$
	Control	0	12	
6	Drug	1	10	$\infty$
	Control	0	10	
7	Drug	1	4	2.0
	Control	1	8	
8	Drug	4	2	0.3
	Control	6	1	

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

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

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;
```

with output:

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  $u_1, \dots, u_8$  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 significant?

## 12.4: Clustered multinomial responses: examples

### 4. Insomnia study

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.

Treatment	Time to falling asleep				
	Initial	Follow-up			
		< 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

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

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

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:

$$P(Y_{ij} = 1) = P(Y_{ij} \leq 1)$$

$$P(Y_{ij} = 2) = P(Y_{ij} \leq 2) - P(Y_{ij} \leq 1)$$

$$P(Y_{ij} = 3) = P(Y_{ij} \leq 3) - P(Y_{ij} \leq 2)$$

$$P(Y_{ij} = 4) = 1 - P(Y_{ij} \leq 3)$$

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;

```

with 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

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  $e^{\beta_2}$  for placebo and  $e^{\beta_2+\beta_3}$  for treatment:

Label	Estimate	Standard		DF	t Value	Pr >  t	Alpha	Lower	Upper
		Error							
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 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;
```

gives 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	Standard		Alpha	Confidence Limits		Wald	
			Estimate	Error		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).