Generalized Linear Mixed Modeling and PROC GLIMMIX

Richard Charnigo
Professor of Statistics and Biostatistics
Director of Statistics and Psychometrics Core, CDART
RJCharn2@aol.com

Objectives

First ~80 minutes:

1. Be able to formulate a generalized linear mixed model for longitudinal data involving a categorical and a continuous covariate.

2. Understand how generalized linear mixed modeling differs from logistic regression and linear mixed modeling.

Last ~40 minutes:

3. Be able to use PROC GLIMMIX to fit a generalized linear mixed model for longitudinal data involving a categorical and a continuous covariate.

Motivating example

The Excel file at {www.richardcharnigo.net/glimmix} contains a simulated data set:

Five hundred college freshmen (“ID”) are asked to indicate whether they have consumed marijuana during the past three months (“MJ”).

The students are also assessed on negative urgency; the results are expressed as Z scores (“NegUrg”).

One and two years later (“Time”), most of the students supply updated information on marijuana consumption; however, some students drop out.

Motivating example

Two possible “research questions” are:

i. Is there an association between negative urgency and marijuana use at baseline ?

ii. Does marijuana use tend to change over time and, if so, is that change predicted by negative urgency at baseline ?

We can envisage more complicated and realistic scenarios (e.g., with additional personality variables and/or interventions), but this simple scenario will help us get a hold of generalized linear mixed modeling and PROC GLIMMIX.

Exploratory data analysis

Before pursuing generalized linear mixed (or other statistical) modeling, we are well-advised to engage in exploratory data analysis.

This can alert us to any gross mistakes in the data set, heretofore undetected, which may compromise our work.

This can also suggest a structure for the generalized linear mixed model and help us to anticipate what the results should be.

Exploratory data analysis

Variable	Label	N	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Mean	Std Dev	Skewness
Time	Time	1350	0.00	0.00	1.00	2.00	2.00	0.93	0.81	0.14
NegUrg	NegUrg	1350	-2.68	-0.63	0.04	0.68	3.03	0.03	0.94	0.06
MJ	MJ	1350	0.00	0.00	0.00	1.00	1.00	0.28	0.45	0.98

In this example, no gross mistakes are apparent. Having Z scores between -2.68 and +3.03 seems reasonable for a sample of size 500. The minimum and maximum values of time and marijuana use are correct, given that the latter is being treated dichotomously.

Exploratory data analysis

Table of MJ by negurgstratum				
MJ(MJ)	negurgstratum			
Frequency Percent Row Pct Col Pct	0	1	2	Total
0	105 21.00 26.79 91.30	207 41.40 52.81 80.54	80 16.00 20.41 62.50	392 78.40
1	10 2.00 9.26 8.70	50 10.00 46.30 19.46	48 9.60 44.44 37.50	108 21.60
Total	115 23.00	257 51.40	128 25.60	500 100.00

There is a clear association between negative urgency stratum and marijuana use during freshman year, with 8.7% of those low on negative urgency (bottom 25%) using marijuana versus 19.5% for average (middle 50%) and 37.5% for high (top 25%).

Exploratory data analysis

Table of MJ by negurgstratum

MJ(MJ)	negurgstratum			
Frequency Percent Row Pct Col Pct	0	1	2	Total
0	94 20.89 30.03 93.07	161 35.78 51.44 68.80	58 12.89 18.53 50.43	313 69.56
1	7 1.56 5.11 6.93	73 16.22 53.28 31.20	57 12.67 41.61 49.57	137 30.44
Total	101 22.44	234 52.00	115 25.56	450 100.00

A similar phenomenon is observed in sophomore year, but overall marijuana use has increased from 21.6% to 30.4%.

Exploratory data analysis

Table of MJ by negurgstratum				
MJ(MJ)	negurgstratum			
Frequency Percent Row Pct Col Pct	0	1	2	Total
0	81 20.25 30.45 88.04	139 34.75 52.26 65.88	46 11.50 17.29 47.42	266 66.50
1	11 2.75 8.21 11.96	72 18.00 53.73 34.12	51 12.75 38.06 52.58	134 33.50
Total	92 23.00	211 52.75	97 24.25	400 100.00

By junior year, marijuana use has increased to 33.5%.

First generalized linear mixed model

Let Y_{jk} denote subject j's marijuana use at time k.
Because Y_{jk} is dichotomous, we cannot employ a linear mixed model, which assumes a continuous (in fact, normally distributed) outcome.

However, consider these three equations:

$\text{logit}\{P(Y_{jk} = 1)\} = a_0 + a_1 k$, if subject j is low

$\text{logit}\{P(Y_{jk} = 1)\} = b_0 + b_1 k$, if subject j is average

$\text{logit}\{P(Y_{jk} = 1)\} = c_0 + c_1 k$, if subject j is high on negative urgency,

where logit{x} is defined as $\log(x / (1-x))$.

First generalized linear mixed model

Three comments are in order:

First, the generalized linear mixed model defined by the three equations can be expressed as a logistic regression model. Let X_1 and X_2 respectively be dummy variables for low and high negative urgency. Then we may write

$$\text{logit}\{P(Y_{jk} = 1)\} = b_0 + (a_0 - b_0) X_{1j} + (c_0 - b_0) X_{2j} + (b_1 + (a_1 - b_1) X_{1j} + (c_1 - b_1) X_{2j}) k.$$

Indeed, just as linear regression is a special case of linear mixed modeling, logistic regression is a special case of generalized linear mixed modeling.

First generalized linear mixed model

Second, we are in essence logistic-regressing marijuana use on time but allowing each subject to have one of three intercepts and one of three slopes, according to his/her negative urgency.

Third, our research questions amount to asking whether a_0 , b_0 , c_0 differ from each other, whether a_1 , b_1 , c_1 differ from zero, and whether a_1 , b_1 , c_1 differ from each other.

First generalized linear mixed model

Now let us examine the results from fitting the generalized linear mixed model using PROC GLIMMIX.

We see that PROC GLIMMIX used all available observations (1350), including observations from the 100 subjects who dropped out early.

Number of Observations Read	1350
Number of Observations Used	1350

First generalized linear mixed model

The estimates of the intercepts a_0 , b_0 , c_0 are -2.48, -1.33, and -0.45. The estimates of the slopes a_1, b_1 , c_1 are 0.18, 0.38, and 0.31. Exponentiating the latter gives us estimates of the factors by which the odds of marijuana use get multiplied each year, within each of the negative urgency strata. For example, exp(0.3143) = 1.369 in the high stratum.

Parameter Estimates						
Effect	negurgstratum	Estimate	Standard Error	DF	t Value	Pr > |t|
negurgstratum	0	-2.4795	0.3179	1344	-7.80	<.0001
negurgstratum	1	-1.3264	0.1386	1344	-9.57	<.0001
negurgstratum	2	-0.4531	0.1668	1344	-2.72	0.0067
Time*negurgstratum	0	0.1815	0.2424	1344	0.75	0.4542
Time*negurgstratum	1	0.3750	0.1049	1344	3.58	0.0004
Time*negurgstratum	2	0.3143	0.1361	1344	2.31	0.0211

First generalized linear mixed model

We can also use PROC GLIMMIX to estimate any linear combinations of a_0 , b_0 , c_0 , a_1 , b_1 , c_1. For example, below are estimates of

$c_0 - a_0$
(high vs. low negative urgency freshmen),
$(c_0 + c_1) - (a_0 + a_1)$
(high vs. low negative urgency sophomores), and
$(c_0 + 2c_1) - (a_0 + 2a_1)$
(high vs. low negative urgency juniors).

Again, exponentiation will yield estimated odds ratios.

Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr > |t|
High vs low freshman	2.0264	0.3590	1344	5.64	<.0001
High vs low sophomore	2.1592	0.2270	1344	9.51	<.0001
High vs low junior	2.2919	0.3589	1344	6.39	<.0001

Second generalized linear mixed model

As noted earlier, our first generalized linear mixed model can be expressed as a logistic regression model. How, then, does generalized linear mixed modeling go beyond logistic regression ?

The answer is that we may also allow each subject to have his/her own personal intercept and slope, not merely choose from among three intercepts and three slopes. This can capture correlations among repeated measurements on that subject. The personal intercept and slope may be related to negative urgency and to unmeasured factors. For simplicity in what follows, however, we will confine attention to a personal intercept.

Second generalized linear mixed model

More specifically, we propose the following:

$$\text{logit}\{P(Y_{jk} = 1)\} = b_0 + (a_0 - b_0)\, X_{1j} + (c_0 - b_0)\, X_{2j} + P_{1j} + (\, b_1 + (a_1 - b_1)\, X_{1j} + (c_1 - b_1)\, X_{2j}\,)\, k.$$

Above, P_{1j} is an unobserved zero-mean variable that adjusts the intercept for subject j. Thus, the interpretations of a_0 , b_0 , c_0 are subtly altered. They are now the average intercepts for subjects who are low, average, and high on negative urgency.

Even so, our research questions are still addressed by estimating a_0 , b_0 , c_0 , a_1 , b_1 , c_1.

Second generalized linear mixed model

While we can "predict" P_{1j} from the data, in practice this is rarely done. However, its variance is routinely estimated.

Covariance Parameter Estimates			
Cov Parm	Subject	Estimate	Standard Error
UN(1,1)	ID	1.2456	0.3290

Solutions for Fixed Effects						
Effect	negurgstratum	Estimate	Standard Error	DF	t Value	Pr > |t|
negurgstratum	0	-2.9656	0.3689	847	-8.04	<.0001
negurgstratum	1	-1.6720	0.1859	847	-8.99	<.0001
negurgstratum	2	-0.5848	0.2156	847	-2.71	0.0068
Time*negurgstratum	0	0.2225	0.2538	847	0.88	0.3810
Time*negurgstratum	1	0.4632	0.1193	847	3.88	0.0001
Time*negurgstratum	2	0.4116	0.1576	847	2.61	0.0092

Second generalized linear mixed model

Some care is now required in interpreting odds ratio estimates. For example, exp(2.3808) = 10.81 says that a freshman high on negative urgency is estimated to have 10.81 times the odds of using marijuana versus a freshman low on negative urgency, controlling for whatever unmeasured factors contribute to the personal intercepts.

Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr > |t|
High vs low freshman	2.3808	0.4195	847	5.68	<.0001
High vs low sophomore	2.5699	0.3046	847	8.44	<.0001
High vs low junior	2.7590	0.4327	847	6.38	<.0001

Second generalized linear mixed model

Which model is better: the first or second ?

Conceptually, the second model is appealing because P_{1j} captures correlations among the repeated observations on subject j. Thus, we avoid the unrealistic assumption, present in logistic regression, that observations are independent.

Empirically, we may examine a model selection criterion such as the BIC; a smaller value is better. Here are results for the first and second models.

Fit Statistics	
-2 Log Likelihood	1467.29
AIC (smaller is better)	1479.29
AICC (smaller is better)	1479.35
BIC (smaller is better)	1510.54

Fit Statistics	
-2 Log Likelihood	1432.21
AIC (smaller is better)	1446.21
AICC (smaller is better)	1446.29
BIC (smaller is better)	1475.71

Third generalized linear mixed model

So far we have treated negative urgency as categorical, but this is not necessary and perhaps not optimal. Let us now consider the following:

$$\text{logit}\{P(Y_{jk} = 1)\} = (d_0 + e_0 N_j + P_{1j}) + (d_1 + e_1 N_j) k.$$

Above, N_j denotes the continuous negative urgency variable, while P_{1j} is, as before, an adjustment to the intercept.

Third generalized linear mixed model

Since negative urgency was expressed as a Z score, d_0 is the average intercept and d_1 is the slope among those average on negative urgency.

Likewise, $d_0 + e_0$ is the average intercept and $d_1 + e_1$ is the slope among those one standard deviation above average on negative urgency.

And, $d_0 – e_0$ is the average intercept and $d_1 – e_1$ is the slope among those one standard deviation below average on negative urgency.

Third generalized linear mixed model

We estimate the variance of P_{1j} as well as estimating d_0 , e_0 , d_1 , e_1 .

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
UN(1,1)	ID	0.9766	0.2923

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-1.7364	0.1501	498	-11.56	<.0001
NegUrg	1.0781	0.1500	848	7.19	<.0001
Time	0.4203	0.09476	848	4.44	<.0001
NegUrg*Time	0.02515	0.1052	848	0.24	0.8111

Third generalized linear mixed model

In addition, we may estimate linear combinations of d_0 , e_0 , d_1 , e_1 . For example, $2e_0$ compares freshmen one standard deviation above to freshmen one standard deviation below, $2e_0 + 2e_1$ compares such sophomores, and $2e_0 + 4e_1$ compares such juniors. Moreover, the BIC prefers this model over either of the first two.

Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr > |t|
High vs low freshman	2.1562	0.2999	848	7.19	<.0001
High vs low sophomore	2.2065	0.2195	848	10.05	<.0001
High vs low junior	2.2569	0.3080	848	7.33	<.0001

Fit Statistics	
-2 Log Likelihood	1391.30
AIC (smaller is better)	1401.30
AICC (smaller is better)	1401.35
BIC (smaller is better)	1422.38

What's next ?

Now we will launch SAS and examine the PROC GLIMMIX implementations of the three generalized linear mixed models.

In addition, although beyond the scope of today's presentation, I mention that there are versions of generalized linear mixed models that accommodate responses which are neither normally distributed nor dichotomous. The most common is a version that accommodates responses which are Poisson distributed; this is useful when the outcome of interest is a count (e.g., how many times a person engages in a particular behavior).