Using PROC NLMIXED and PROC GLMMIX to analyze dyadic data with binary outcomes

Peter L. Flom, National Development and Research Institutes, New York, NY
James M. McMahon, National Development and Research Institutes, New York, NY
Enrique R. Pouget , National Development and Research Institutes, New York, NY

ABSTRACT

In the social and health sciences, data are often hierarchical (subjects nested in groups). One kind of hierarchy is the dyad, or couple, where each group consists of two subjects. Dyadic data pose particular problems for statistical analysis for several reasons: First, variation may occur at the individual or dyadic level. Second, the data are not independent. Third, the small group size poses special difficulties. Multilevel models have been used for dyadic data; we demonstrate the use of PROC NLMIXED and PROC GLIMMIX, and discuss the strengths and weaknesses of this approach in general, and these SAS procedures in particular. We illustrate this with data on predictors of viral Hepatitis C among heterosexual couples in Harlem in New York City.

Keywords: Dyadic Multilevel NLMMIXED GLIMMIX binary.

INTRODUCTION

The most commonly used statistical techniques (e.g. the general linear model) assume that the data are independent. For data that come from individuals, this often makes sense — what one subject says is often unrelated to what other subjects say. However, many data sets are hierarchical, that is, the data are nested. Two common kinds of hierarchy are temporal and spatial clustering; one commonly cited example of spatially clustered data is students in classrooms in schools; temporal clustering usually involves repeated measures on a subject. In this paper, we discuss another type of clustering, one involving couples, or *dyads*; a dyad is made up of an *actor* and a *partner*. The actors are the people who respond to the questions, and the partners are the people who they are in relationships with. Our particular example involves behaviors among heterosexual couples. Here, we assume that the dyads are independent; other methods must be used when this is not the case. However, the relationships that make the data dyadic violate the assumption of independence. The methods we discuss deal with this violation of independence. The methods we propose are multilevel models (also known as hierarchical linear models, mixed models, and various other terms). Several other models have been proposed for such data, but space does not permit us to discuss them; for a review see Campbell & Kashy (2002) and references cited there.

In the following sections we

- Introduce the example
- Introduce the multilevel approach
- Detail its assumptions
- Discuss types of dyadic variables and data structure
- Discuss coding and centering
- Illustrate the use of PROC NLMIXED and PROC GLIMMIX
- Discuss the results
- Summarize the paper

EXAMPLE: HEPATITIS C AMONG HETEROSEXUAL COUPLES

Hepatitis C is the most common chronic blood-borne infectious disease in the US and is a major cause of morbidity and mortality nationwide; nearly 4 million people in the US are infected (Alter 1999). Past and current drug users constitute the largest group of persons infected with the virus in the US, and the vast majority of new infections occur in drug injectors (Edlin 2002). Studies of injectors often report prevalence of 75% or more (Lorvick et al. 2001, McCarthy & Flynn 2001, Stein et al. 2001) and incidences of between 10% and 20% per year (Hagan et al. 1999, Garfein et al. 1998, Thorpe et al. 2002).

We report findings from a study of risk for Hepatitis C and other infections among drug-using, heterosexual couples in East Harlem, a low-income, mainly African-American and Latino neighborhood of New York City (Tortu et al. 2003, McMahon

et al. 2003). A total of 265 couples with conclusive Hepatitis C Virus (HCV) test results for each partner were enrolled; the dependent variable was actor HCV antibody reactivity, and independent variables included actor gender (aGen), actor injection drug use (aIDU), partner age (pAge) and recent dyadic sexual behavior (dSex). These variables were selected to help illustrate the methods; the models should not be viewed substantively. A more complex analysis will be presented in future work.

TYPES OF DYADIC VARIABLES AND DATA STRUCTURE

Discussion of dyadic data requires some definitions that may be unfamiliar. There are three types of independent variables in dyadic data: Within-dyad, between-dyad, and mixed. A between-dyad variable is one that varies across dyads, but where both members of the dyad should report the same thing, although there may be measurement error (e.g. length of relationship); a within-dyad variable is one that varies within the dyad, but not across dyads (e.g., in our study, gender is a within dyad variable since each couple has one man and one woman); and finally, a mixed variable is one that can vary both within and across dyads (e.g. most demographic variables are mixed). Actor and partner effects can only be estimated for mixed variables, or for interactions that involve mixed variables (Campbell & Kashy 2002).

Data must be structured in a particular way for SAS PROCs to be able to use it. In particular, each row must refer to a person, not a dyad, and there must be a variable indicating dyad membership and, for those cases where dyad members are distinguishable, there must be a variable indicating that (e.g., for heterosexual couples, this variable would be 'male' or 'female').

MULTILEVEL MODELING APPROACHES TO DYADIC DATA WITH BINARY OUTCOMES

Multilevel modeling was developed to deal with dependent data. We go over it briefly here, several recent books have dealt with it, including Raudenbush & Bryk (2002).

The general linear model can be represented, in matrix terms, as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{1}$$

where $\mathbf{Y}$ is a vector of response or dependent variables, $\mathbf{X}$ is a matrix of independent variables or covariates, β is a vector of parameters to be estimated, and ε is a vector of errors. One assumption of this model is that

$$\mathbf{e} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

In our example, both the assumption of independence and that of normality are problematic. Independence is violated because the data are dyadic — characteristics such as health, attitudes, and behavior of one partner are likely to be related to characteristics of the other. The normality assumption is violated because the dependent variable is a dichotomy (positive or negative for HCV antibody). Therefore, two generalizations of this model are necessary: One to deal with the dependence, and one to deal with the non-normality. In addition, the dyadic nature of the data poses special problems. We discuss each of these problems below.

THE DEPENDENCE PROBLEM

The dependency problem can be accommodated by including both fixed and random effects in the model. Fixed effects are those which assume no sampling error or random variance (e.g. group means); random effects are those that estimate population variance and include sampling variation.

Perhaps the simplest way to describe the models we will be using is to describe them as two-level models, with level one being at the level of the individual, and level two being at the level of the dyad. Because the data are dyadic, we will only be able to include a single random effect, that is, we will use random intercepts for each dyad, but assume fixed slopes. For this case, if we assume a single independent variable then the models are

Individual level model

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \varepsilon_{ij} \tag{2}$$

which is an ordinary regression equation about the response of the ith individual in the jth dyad and

Dyad level model

$$\beta_{0j} = \gamma_{00} + \nu_{0j} \tag{3}$$

$$\beta_{1j} = \gamma_{10} \tag{4}$$

saying that the intercept (β_{0j}) in equation 2 is related to a general level (γ_{00}) and a component specific to each dyad (ν_{0j}) but all the slopes are the same. Here, γ_{00} andγ_{10} are fixed effects, and ν_{0j} is a random effect.

These models can be fit with SAS PROC MIXED.

THE DYADIC DATA PROBLEM

Dyadic data poses particular problems because each group is so small (n = 2). One method of dealing with this problem is the Actor-Partner Interdependence Model (APIM) Kashy & Kenny (2000). Campbell & Kashy (2002) illustrated how to implement this model with PROC MIXED. First, the data must be structured correctly, that is, there must be one observation for each person, or two observations for each dyad; for cases with distinugishable dyads, there must also be a variable indicating which part of the dyad the subject is (e.g., male or female). Second, categorical variables should be coded using effect or dummy coding, (Campbell & Kashy 2002, McMahon et al. 2006). Third, the quantitative independent variables should be centered around their grand means, again for ease of interpretation (Campbell & Kashy 2002). Fourth, any interactions that you wish to include in the model should be coded in a DATA step. Following this, PROC MIXED can be run; some sample code, adopted from Campbell & Kashy (2002) is:

```
proc mixed data = new;
    class id;
    model wdraw =  asecure psecure agen cond / solution ddfm = satterth;
    random intercept/type = cs subject = id;
run;
```

The `solution` option requests the parameter estimates from the independent variables. The `ddfm = satterth` option requests the Sattherthwaite approximation to the denominator degrees of freedom. The `random` statement treats the individual scores as random effects, and the `type = cs` option tells SAS to use a compound symmetry structure for the non-independence of the dyad members.

Specific problems posed by dyadic data preclude the estimation of both random slopes and intercepts which would result in an overdetermined model (Newsom 2002), and can bias the estimates and confidence limits for the dyadic level random variance components can be biased because of the small number of subjects per cluster (Hox 1998, Hox & Maas 2002, Raudenbush & Bryk 2002).

THE DICHOTOMOUS DEPENDENT VARIABLE PROBLEM

The final problem in our example is that the dependent variable is dichotomous. If the data were independent, we could deal with this by using PROC LOGISTIC or PROC GENMOD. For mixed models, SAS supplies two procedures: PROC NLMIXED and PROC GLIMMIX, details of which are presented below. Here we discuss some complications that dichotomous variables present for multilevel models.

Logistic regression transforms the dependent variable using the logit link:

$$\eta_{ij} = log\left(\frac{p_{ij}}{1-p_{ij}}\right)$$

For multilevel models, the link yields the following:

Individual level model

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} \tag{5}$$

Notice that here there is no error term, this is because the the variance of the dependent variable can be derived directly from the dependent variable (DV) itself, and is therefore considered fixed (Snijders & Bosker 1999, Raudenbush & Bryk 2002).

Dyad level model

$$\beta_{0j} = \gamma_{00} + \nu_{0j} \tag{6}$$
$$\beta_{1j} = \gamma_{10} \tag{7}$$

which is identical to the level two model for a normally distributed DV.

These models can be combined via simple algebra to

$$\eta_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \nu_{0j} \tag{8}$$

In multilevel models with normally distributed DVs, the next step is often to examine the intraclass correlation coefficient (that is, the proportion of variance of an observation due to between-dyad variability (Everitt 1998)). Some recommend that, if this

is not significantly different from 0, then the multilevel model is not needed. There are two problems with this approach, one of which is particular to dyadic DVs, and the other more general. The general problem is that it is not the significance of the dependence which should matter, but its practical importance and theoretical significance; in this particular case, if what one member of a heterosexual couple said about their relationship was independent of what the other one said, that would be either a sign of serious data problems or a highly important finding in its own right. The more particular problem is that there is no exact equivalent of the ICC for logistic models (just as there is no exact equivalent of R^2 for logistic regression), nor is there agreement as to what level of significance is 'enough', with Snijders & Bosker (1999) arguing for a p-value of 0.20 or 0.25. These and other authors recommend using the pairwise correlations in place of the ICC to assess dyadic dependence when the outcome is binary.

SAS PROCS FOR MULTILEVEL MODELS

There are two SAS PROCs that analyze nonlinear mixed models: PROC NLMIXED and PROC GLIMMIX. The latter is available only in v 9, and must be downloaded from the SAS website. We briefly discuss the two here, in a relatively nontechnical way. For more information, see the SAS documentation.

ASSUMPTIONS OF THE MODEL

This model assumes that (a) The probability of 'success' ($y_{ij} = 1$) is the same for both individuals in a dyad; (b) Observations between clusters are independent, and those within clusters have identical correlations; (c) Each random effect is independent, and each has a distribution that can be estimated via maximum likelihood; (d) Random effects and model predictors at all levels are independent; (e) There is some appropriate model linking y_i and u_i, and it has some joint probability density function (for a fuller discussion of these issues, see Raudenbush & Bryk (2002)).

PROC NLMIXED

PROC NLMIXED was introduced in version 7 of SAS. It produces likelihood estimates that are maximized exactly in theory, and based on adaptive Gaussian quadrature (Pinheiro & Bates 2000), and can handle a wide variety of dependent variables — indeed, it allows one to program one's own distribution if it is not provided (Gaussian quadrature is a method for preforming numerical integration using a series expansion of the form

$$\int f(x)\phi(x)dx \approx \sum_{i=1}^{m} w_m f(x_m)$$

where x_m are the Gaussian quadrature points and w_m are the weights; these are available from tables Everitt (1998).

NLMIXED code McMahon and his colleagues (McMahon et al. 2006) used the following SAS code:

```
proc nlmixed data=couplesHCV qpoints=20 tech=newrap ;
 title 'Example NLMIXED analysis with HCV couples data';
 parms beta0= -1.793 beta1=-0.176 beta2=3.054 beta3=0.056 beta4=0.016 s2u=0.042;
  eta = beta0 +beta1*aGEN + beta2*aIDU + beta3*pAGEc + beta4*dSEXc + u;
  mu = exp(eta) / (1 + exp (eta));
  model aHCV ~ binary (mu);
  random u ~normal(0, s2u) subject=id;
run;
```

`QPOINTS` sets the number of quadrature points, Carlin et al. (2001) recommend at least 20. `TECH` sets the optimization method, Newton-Raphson is time-intensive, but tends to be among the most reliable methods (SAS Institute Inc. 1999). `PARMS` sets the parameter starting values, which can be estimated using other SAS PROCs (e.g. here, GENMOD and MIXED (McMahon et al. 2006)). `ETA` specifies the combined multilevel model. `MU` specifies the link function (here logistic), `MODEL` specifies the DV and its distribution, and specifies the model, `RANDOM` identifies the random effects and their distribution, and, finally, the `subset = id` statement identifies dyad membership.

NLMIXED results for random intercepts model

The NLMIXED Procedure

Specifications

Data Set	WORK.COUPLESHCV
Dependent Variable	aHCV
Distribution for Dependent Variable	Binary
Random Effects	u
Distribution for Random Effects	Normal
Subject Variable	id
Optimization Technique	Newton-Raphson
Integration Method	Adaptive Gaussian Quadrature

Dimensions

Observations Used	530
Observations Not Used	0
Total Observations	530
Subjects	265
Max Obs Per Subject	2
Parameters	6
Quadrature Points	20

Parameters

beta0	beta1	beta2	beta3	beta4	s2u	NegLogLike
-1.793	-0.176	3.054	0.056	0.016	0.042	264.206538

These tables are useful for checking that SAS did what you think you told it to do.

Iteration History

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	16	260.347216	3.859322	11.62197	-6.06859
2	24	258.910991	1.436225	2.753375	-2.32362
3	32	258.573991	0.337	0.582256	-0.57958
4	40	258.545085	0.028905	0.075222	-0.05425
5	48	258.544772	0.000314	0.001022	-0.00062
6	56	258.544772	4.536E-8	6.001E-8	-9.07E-8

NOTE: GCONV convergence criterion satisfied.

The iteration history shows that the procedure converged. This can be useful for diagnosing problems.

Fit Statistics

-2 Log Likelihood	517.1
AIC (smaller is better)	529.1
AICC (smaller is better)	529.3
BIC (smaller is better)	550.6

The fit statistics are primarily useful for comparing one model to another

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
beta0	-2.3452	0.3747	264	-6.26	<.0001	0.05	-3.0830	-1.6075	2.396E-8
beta1	-0.2186	0.2515	264	-0.87	0.3855	0.05	-0.7139	0.2766	9.475E-9
beta2	4.0417	0.5053	264	8.00	<.0001	0.05	3.0466	5.0367	5.277E-8
beta3	0.06589	0.0204	264	3.23	0.0014	0.05	0.02574	0.1060	-5.84E-8
beta4	0.02027	0.0089	264	2.27	0.0238	0.05	0.002709	0.03784	-4.29E-8
s2u	2.0300	0.9358	264	2.17	0.0310	0.05	0.1874	3.8727	-6E-8

These are the key results. The table lists the six free parameters, their maximum likelihood estimates, standard errors, and inferential statistics. β_0 is the intercept, and represents the log-odds of anti-HCV reactivity for a person with 0 on all the other variables. To convert the log-odds back to probabilities, use the inverse exponential function:

$$P_{ij} = \frac{e^{\eta_{ij}}}{1 - e^{\eta_{ij}}}$$

The next rows of output are coefficient estimates for aGEN (β_1), aIDU (β_2), pAGEc (β_3), dSEXc (β_4), and the level 2 random effect (s_u^2). Each can be converted to an adjusted odds ratio by exponentiating it. The results indicate that injection drug use, partner age, and recent unprotected sex are significant predictors of actor HCV status. The first two are individual level predictors, while the last is a dyadic level predictor. Once again, we emphasize that this model is for illustrative purposes only, and should not be taken to represent substantive findings.

PROC GLIMMIX

PROC GLIMMIX has some advantages compared to NLMIXED, but also some disadvantages. From a practical standpoint, one of the big advantages is that the syntax is very similar to PROC MIXED, and somewhat similar to PROC GLM, and is therefore likely to be more familiar. Statistically, it can handle more random effects than NLMIXED. Also, in version 9.1, ODS graphics are available for GLIMMIX, but not for NLMIXED. Disadvantages of GLIMMIX are that the dependent variable has to be from an exponential distribution, whereas NLMIXED allows more flexibility (e.g. it can fit zero-inflated models), and that NLMIXED offers a true log likelihood, which GLIMMIX does not. A more detailed comparison of the two procedures is given below.

GLIMMIX code

```
proc glimmix data = couplesHCV ;
  title 'Example GLIMMIX code with HCV couples data';
  model aHCV = aGEN aIDU pAGEc dSEXc/solution link = logit dist = binomial;
  random intercept /subject = id gcorr ;
   run;
```

GLIMMIX results for random intercepts model

Model Information

Data Set	WORK.COUPLESHCV
Response Variable	aHCV
Response Distribution	Binomial
Link Function	Logit
Variance Function	Default
Variance Matrix Blocked By	id
Estimation Technique	Residual PL
Degrees of Freedom Method	Containment
Number of Observations Read	530
Number of Observations Used	530

Dimensions

G-side Cov. Parameters	1
Columns in X	5

Columns in Z per Subject	1
Subjects (Blocks in V)	265
Max Obs per Subject	2

Optimization Information

Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	1
Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Profiled
Starting From	Data

Iteration History

Iteration	Restarts	Subiterations	Objective Function	Change	Max Gradient
0	0	4	2450.2692821	0.80937162	2.163E-7
1	0	3	2473.0874466	0.09068533	0.000044
2	0	2	2476.4848549	0.02793991	0.00004
3	0	3	2476.5400813	0.00530558	4.934E-8
4	0	1	2476.5465449	0.00099212	2.427E-6
5	0	1	2476.5469045	0.00018201	8.324E-8
6	0	1	2476.5469396	0.00003329	4.072E-9
7	0	1	2476.5469449	0.00000608	7.612E-8
8	0	1	2476.5469458	0.00000235	0.000018
9	0	1	2476.5469462	0.00000169	0.000013
10	0	0	2476.5469459	0.00000000	8.45E-6

Convergence criterion (PCONV=1.11022E-8) satisfied.

Fit Statistics

-2 Res Log Pseudo-Likelihood	2476.55
Generalized Chi-Square	416.26
Gener. Chi-Square / DF	0.79

These have similar interpretations to the output from NLMIXED.

Estimated G Correlation Matrix

Effect	Row	Col1
Intercept	1	1.0000

Because there is only one random effect (the intercept) this table is not particularly useful.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	id	0.6953	0.3083

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-1.8006	0.2522	263	-7.14	<.0001

aGEN	-0.1819	0.2202	262	-0.83	0.4095
aIDU	3.0956	0.2758	262	11.22	<.0001
pAGEc	0.05932	0.01664	262	3.56	0.0004
dSEXc	0.01647	0.006925	262	2.38	0.0181

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
aGEN	1	262	0.68	0.4095
aIDU	1	262	125.94	<.0001
pAGEc	1	262	12.71	0.0004
dSEXc	1	262	5.65	0.0181

The parameter estimates can be interpreted in a similar way as those for NLMIXED. Note that the variables are now named, and that the covariance parameter estimate of 0.695 is the equivalent here of s_u^2 in the NLMIXED table. It can be seen that, although the parameter estimates generated by NLMIXED and GLIMMIX are slightly different (and very different for the random effect) the conclusions are the same for both methods.

COMPARING NLMIXED AND GLIMMIX

There are several important differences between NLMIXED and GLIMMIX that analysts should consider in choosing a procedure. The primary difference lies in the estimation method used by each procedure. Both procedures approach parameter estimation as an optimization problem, which solves for an approximation of the marginal log likelihood. NLMIXED accomplishes this using an integral approximation through Gaussian quadrature, whereas GLIMMIX relies on approximation of a linear mixed model (linearization). Each method has a number of advantages and disadvantages. Advantages of the NLMIXED method are that it is generally more accurate and generates a true log-likelihood fit statistic that can be used to compare nested models. The method also permits greater flexibility to accommodate user-defined likelihood functions. GLIMMIX, in contrast, can produce potentially biased estimates for both fixed effects and covariance parameters, especially for binary data (Schabenberger 2005). GLIMMIX generates Wald-type test statistics and confidence intervals and nested models cannot be compared with true likelihood ratio tests. The trade-off for less accurate estimates in GLIMMIX is that it allows greater flexibility in the types of models that can be estimated, the number of random effects that can be specified, and the fit options available. For example, GLIMMIX allows multiple nested and crossed random effects, whereas NLMIXED cannot accommodate a large number of random effects (< 5) and is limited to only two levels. Additionally, GLIMMIX allows the use of restricted maximum likelihood (REML) methods, which have been shown to produce better estimates than full maximum likelihood (ML) when the number of higher-level units is small. REML is not available in NLMIXED. GLIMMIX also supports model-based and sandwich estimation for standard errors, whereas NLMIXED supports only model-based standard errors. Sandwich estimation provides consistent results even if the variance function is misspecified Everitt (1998). However, as noted above, GLIMMIX requires that the dependent variable be from an exponential family, whereas NLMIXED allows the user to write his or her own function (e.g. zero-inflated models).

A further difference between the two procedures is in the way initial parameter values are generated and applied. For NLMIXED, the user must generate parameter starting values and enter these values into the SAS code prior to running the procedure. Generally, other SAS procedures such as PROC MIXED or PROC GENMOD are used to generate the initial values for NLMIXED. In contrast, GLIMMIX uses a double iteration scheme in which parameter starting values are generated from an iteratively-derived approximated linear mixed model. These initial values are then used to update the linearization, and are then applied to iteratively fit a second final linear mixed model. While the GLIMMIX approach requires less effort on the part of the user, the lack of user control can be disadvantageous in certain situations. Final parameter estimates are generally quite sensitive to modifications of the initial values and convergence can fail due to ill-defined values. By allowing the user to define the initial starting values, NLMIXED provides an opportunity to weigh this sensitivity and modify values to allow convergence.

There are two basic methods for handling correlated data: one is using a marginal methods approach such as the generalized estimating equation (GEE) (Liang & Zeger 1986, Zeger & Liang 1986), which does not incorporate random effects but simply models the correlation in the data—the so-called R-side covariance method. Another is to specify a mixed model incorporating random effects—the G-side random effects method. Because GLIMMIX relies on the linearization estimation technique it can implement either R-side or G-side estimation methods, whereas NLMIXED is limited to G-side random effects estimation.

In summary, NLMIXED is appropriate for nonlinear mixed model estimation when the models are simple and limited to two levels, the number of random effects is small, and the number of level-2 groups is relatively large. Given these conditions,

NLMIXED is recommended for analysis of binary data that require accurate covariance parameter estimates and for models that require user-defined response distributions, or cases in which nested models need to be compared using the likelihood ratio test. Murray et al. (2004) further suggest that the numerical integration maximum likelihood estimation method employed by NLMIXED is superior for multilevel analysis involving small groups, such as family studies. Thus, NLMIXED may be more appropriate for analysis of dyadic data. GLIMMIX is recommended for more complex models, models with a large number of random effects or more than two levels, and models in which the number of higher-level units is small.

DISCUSSION

The results from GLIMMIX and NLMIXED largely agree, which should give us more confidence in both; this is especially notable because dyadic data with dichotomous dependent variables can be very hard to analyze. As noted above, we do not intend this model to be taken as final, and believe that other terms play an important role; again as noted, this paper was intended more to illustrate the methods than to make substantive findings.

SUMMARY

Although dyadic data with dichotomous dependent variables pose some significant problems to data analysis, there is nonetheless much that can be done. While for these data, PROCs GLIMMIX and NLMIXED gave similar results, theory suggests that PROC NLMIXED may be superior for dyadic data in general.

REFERENCES

Alter, M. J. (1999), 'Hepatitis C virus infection in the united states', *Journal of Hepatology* **31**(Suppl 1), 556–562.

Campbell, L. & Kashy, D. A. (2002), 'Estimating actor, partner, and interaction effects for dyadic data using PROC MIXED and HLM: A user-friendly guide', *Personal Relationships* **9**, 327–342.

Carlin, J. B., WOlfe, R. & C. H. Brown, A. G. (2001), 'A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes', *Biostatitistics* **2**, 397–416.

Edlin, B. R. (2002), 'Prevention and treatment of hepatitis C in injection drug users', *Hepatology* **36**, S210–S219.

Everitt, B. S. (1998), *The Cambridge dictionary of statistics*, Cambridge University Press, Cambridge, UK.

Garfein, R. S., Doherty, M. C., Monterroso, E. R., Thomas, D. L., Nelson, K. E. & Vlahov, D. (1998), 'Prevalence and incidence of hepatitis C virus infection among adult injection drug users', *Journal of Acquired Immune Deficiency Syndromes & Human Retrovirology* **18**(supplement 1), S11–S19.

Hagan, H., McGough, J. P., Thiede, H., Weiss, N. S., Hopkins, S. & Alexander, E. R. (1999), 'Syringe exchange and risk of infection with hepatitis B and C viruses', *American Journal of Epidemiology* **149**(3), 203–213.

Hox, J. J. (1998), *Multilevel modeling: When and Why? Classification, data analysis and data highways*, Springer, Berlin.

Hox, J. J. & Maas, C. J. M. (2002), Sample sizes for multilevel modeling, *in* 'Social science methodology in the new millenium', Fifth international conference on logic and methodology.

Kashy, D. A. & Kenny, D. A. (2000), The analysis of data from dyads and groups, *in* H. T. Rice & C. M. Judd, eds, 'Handbook of Research Methods in Social Psychology', Cambridge University Press, New York, pp. 451–457.

Liang, K. Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22.

Lorvick, J., Kral, A. H., Seal, K. H., Gee, L. & Edlin, B. R. (2001), 'Prevalence and duration of hepatitis C among injection drug users in san francisco', *American Journal of Public Health* **91**, 46–47.

McCarthy, J. J. & Flynn, N. (2001), 'Hepatitis C in methadone maintenance patients', *Journal of Addictive Diseases* **20**, 19–31.

McMahon, J. R., Pouget, E. R. & Tortu, S. (2006), 'A guide for multilevel modeling of dyadic data with binary outcomes using SAS PROC MIXED', *Computational Statistics and Data Analysis* .

McMahon, J. R., Tortu, S., Torres, L., Pouget, E. R. & Hamid, R. (2003), 'Recruitment of heterosexual couples in public health research', *BMC Medical Research Methodology* **4**, 1–12.

Murray, D. M., Varnell, S. P. & Blitstein, J. L. (2004), 'Design and analysis of group randomized trials: A review of recent methodological developments', *American Journal of Public Health* **94**, 423–432.

Newsom, J. T. (2002), 'A multilevel structural equation model for dyadic data', *Structural equation modeling* **9**, 431–447.

Pinheiro, J. C. & Bates, D. M. (2000), *Mixed-effects models in S and S-Plus*, Springer-Verlag, New York.

Raudenbush, S. W. & Bryk, A. S. (2002), *Hierarchical linear models*, 2nd edn, Sage, Thousand Oaks, CA.

SAS Institute Inc. (1999), *SAS/STAT user's guide, version 8. Volume 1*, SAS Institute Inc., Cary, NC.

Schabenberger, O. (2005), Introducing the GLIMMIX procedure for genearlized linear mixed models, *in* 'SAS Users group 30th annual SAS Users Group International Conference', pp. Paper 196–30.

Snijders, T. A. B. & Bosker, R. J. (1999), *Multilevel analysis: An introduction to basic and advanced multilevel modeling*, Sage, London.

Stein, M. D., Maksad, J. & Clarke, J. (2001), 'Hepatitis C disease among injection drug users: knowledge, perceived risk, and willingness to receive treatment', *Drug and Alcohol Dependence* **61**, 211–215.

Thorpe, L. E., Ouellet, L. J., Levy, J. R., Williams, I. T. & Monterroso, E. R. (2002), 'Hepatitis C virus infection: Prevalence, risk factors, and prevention opportunities among young injection drug users in chicago, 1997-1999', *Journal of Infectious Diseases* **182**(6), 645–653.

Tortu, S., McMahon, J. R., Pouget, E. R. & Hamid, R. (2003), She said, he said: Drug using women's perceived vs. actual HIV risk from primary male partners, San Francisco.

Zeger, S. L. & Liang, K. Y. (1986), 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics* **42**(1), 121–130.

ACKNOWLEDGMENTS

This work was supported by NIDA grant P30DA11041, I would also like to thank Ron Fehd for providing help with LaTeX.

CONTACT INFORMATION

Peter L. Flom
National Development and Research Institutes, Inc.
71 W. 23rd St.
8th floor
New York, NY 10010
Phone: (212) 845-4485
Fax: (917) 438-0894
flom@ndri.org
Personal webpage: `www.peterflom.com`
Work webpage: `http://cduhr.ndri.org`

SAS® and all other SAS Institute Inc., product or service names are registered trademarks ore trademarks of SAS Institute Inc., in the USA and other countries. ® indicates USA registration. Other brand names and product names are registered trademarks or trademarks of their respective companies.