

Paper 197-30

Analyzing Incomplete Binary Repeated Measures Data Using SAS®

Bin Yang, Eli Lilly and Company, Indianapolis, Indiana

ABSTRACT

This paper presents a case study in longitudinal data analysis where the goal is to estimate the efficacy of a new drug for treatment of MDD. Data characteristic indicate:

1. Subjects from different treatment groups drop out differentially across time.
2. There are a high proportion of subjects who never experience any response.

To overcome these challenges, we developed a logistic random-effects model with random intercepts. While the model is specified conditionally on subject random effect variable, we also draw inferences on population-averaged important to the assessment of the treatments' efficacy in a population. Specifically, we present and describe using SAS Proc NLMIXED and %GLIMMIX macro to fit the logistic random-effects model.

Key Words: binary; longitudinal; random effect; Proc NLMIXED; %GLIMMIX.

INTRODUCTION

In a clinical trial, we try to determine if the onset of response or rate of improvement due to Treatment A is faster than that of Treatment B. Therefore, measurements of a subject are taken repeatedly that we can model the process of change within individuals. Repeated measures within a subject are usually positively correlated. The correlational structure of the longitudinal data can be described by random subject effects. Random subject effects indicate the degree of subject variation that exists in the population of subjects.

Data from studies with repeated measurement in general are incomplete due to drop out. We will use terminology of little and Rubin (1987, Chapter 6) for the missing-value process. A non-response process is said to be missing completely a random (MCAR) if the missing is independent of both unobserved and observed data and missing at random (MAR) if, conditional on the observed data, the missing is independent of the unobserved measurements. When either of these is plausible, with a likelihood-based analysis it is not necessary to model the missingness mechanism (Agresti A. 2002 Chapter 12). LOCF analysis may be inappropriate, particularly when subject's dropout for reasons related to the response. Logistic random-effects model with random intercepts and/or random slopes offers a useful approach for analyzing incomplete binary longitudinal clinical trial data. The random-effects method is valid under the less stringent assumption of MAR. Such a full longitudinal approach is also very sensible even when the interest is focusing on the treatment effect at the last scheduled visit. Subjects are not assumed to be measured on the same number of timepoints, thus, it takes all information into account, not only from complete observations, but also from incomplete ones, through the conditional expectation of the missing measurements given the observed ones.

PROC NLMIXED and GLIMMIX macro in SAS can fit Logistic random-effects models. %GLIMMIX macro use *Restricted/residual Pseudo-Likelihood* (REPL). REPL maximizes likelihood of a pseudo variable and is therefore not fully MAR. Proc NLMIXED use *improved maximum likelihood* (ML). The procedure is valid as long as the missing values are MAR.

In general, the logistic random-effects model can be defined as a model which satisfies: Let Y_{ij} be the measurement of subject i , measured at time j , $i=1, \dots, N$, $j=1, \dots, n_i$. Let x_{ij} be the fixed effects and covariates in the logistic regression model and b_i a vector of random effects following some specified distribution, usually a multivariate normal with mean 0 and covariance matrix D.

$$E(Y_{ij} | b_i, x_{ij}) = \pi_{ij}; \quad \text{logit}(\pi_{ij}) = x_{ij}'\beta + z_{ij}'b_i$$

A basic characteristic of this model is the inclusion of random subject effects into logistic regression models in order to account for the influence of subjects on their repeated observations. The design vector

z_{ij} represents the covariates of ij th measurement belonging to the random effects. Elements of β estimate subject-specific (SS) but not population-average (PA) effects. SS and PA have different interpretations and are appropriate in different circumstances (Zeger, Liang and Albert, 1998). The difference between SS and PA grows as between subjects' variation increases.

MATERIALS AND METHODS

DATA

Data presented here come from a clinical trial study in which all patients were assigned to treatment with Treatment A or Treatment B (N = 128 and 139 respectively). The Hamilton Depression Rating Scale (HAMD17) is used to measure the depression status of the patients. The response is defined as 1 if reduction in HAMD17 total Score from baseline is greater than 50% and 0 otherwise. For each patient, a baseline assessment and at least one post-baseline assessment are available. Post-baseline visits range from visit 3 to visit 8.

MODEL FITTING

Our approach utilizes a logistic random-effect model in the form:

$$(\Pr(Y_{ij} = 1 | u_i)) = \beta_0 + u_i + \beta_1 \text{visit}_{ij} + \beta_2 \text{baseline}_i + \beta_4 \text{trt}_i + \beta_5 \text{trt}_i * \text{visit}_{ij}$$

where $u_i \sim \text{Normal}(0, \sigma^2)$

This logistic random-effect model includes a *random intercept* to account the variability between subjects. Fixed categorical effects include treatment, visit, and treatment-by-visit interaction, as well as the continuous fixed covariates of baseline score.

THE NL MIXED PROCEDURE

The NL MIXED procedure provides *improved maximum likelihood* (ML) estimates. Unlike the GENMOD procedure, it allows for the explicit modeling of random effects. The NL MIXED procedure allows Newton-Raphson or Quasi-Newton procedures to maximize the likelihood and adaptive Gaussian quadrature can be used to integrate out the random effects. At least theoretically, it delivers exact ML estimates of the parameters if the number of quadrature points is large enough.

Since the NL MIXED Procedure lacks a class statement, so the user must code dummy variables for a particular treatment design and categorical variables. The primary function of coding dummy variables is to examine individual levels of fixed effects treatments by assigning co-efficient (e.g., 0 or 1) to different level of treatment. In effect, a value of 0 drops that treatment level from analysis, while a value of 1 includes the level of that treatment. We also need 6-1 dummy variables for fixed categorical effects visit.

visit	vt4	vt5	vt6	vt7	vt8
Visit 3	0	0	0	0	0
Visit 4	1	0	0	0	0
Visit 5	0	1	0	0	0
Visit 6	0	0	1	0	0
Visit 7	0	0	0	1	0
Visit 8	0	0	0	0	1

The following SAS code fits the described model by adaptive Gaussain quadrature:

```
proc nlmixed data=mmrminv qpoints=50;
  parms  beta0=1.1153 beta1=-0.5384 beta2=-0.2731
         beta4=1.0309 beta5=0.5709 beta6=0.5871
         beta7=0.9937 beta8=1.6750 beta9=1.7154
         beta10=3.0652 beta11=4.6931 beta12=5.0499
         beta13=4.9504 ;
  eta = u +beta0+beta1*TRT+ beta2*BASCT+ beta4*TRT*vt4+
        beta5*TRT*vt5+beta6*TRT*vt6 +beta7*TRT*vt7+
        beta8*TRT*vt8+beta9*vt4+ beta10*vt5+ beta11*vt6+
        beta12*vt7+ beta13*vt8;
  expeta = exp(eta);
  p = (expeta/(1 + expeta));
  model RES ~ binary(p);

  random u ~ normal(0, sigmau*sigmau) subject=PATIENT;

  estimate 'logit difference at visit 8' beta1+beta8;
  estimate 'odds ration at visit 8' exp(beta1+beta8);
  predict p out=predp;
  ods output ParameterEstimates=ESTIMDATA;
  ods output AdditionalEstimates=ESTIMDIFF;
run;
```

Assign initial values to parameters

Define the logistic model model

Specify the random effect and its distribution

Estimating the quantity of interest

Selected NLMIXED statement:

- PARMS** Lists names of parameters and specifies initial values. Provision of precise initial parameter estimates promotes convergence. Parameters not listed in PARMS statement are assigned an initial value of 1.
- Model** Specifies the conditional distribution of the data given the random effects. The types of models that can be fit include normal, binary, binomial, poisson, gamma, and negative binomial.
- Random** Defines the random effects and their distribution. The only distribution currently available for the random effects is normal (m, v) with mean m and variance v . The subject= patient determines when new realizations of the random effects are assumed to occur. The input data set should be clustered according to this variable. In principle, it's possible to include random slopes as well as random intercepts in any of these models, but (particularly with binary data) the data often provide little information to estimate random slopes.
- Estimate** Enables you to compute an additional estimate that is a function of the parameter values. SAS ignores the random effect if a specific value is not assigned.
- Predict** Enables you to construct predictions of an expression across all of the observations in the input dataset.
- Qpoints** The likelihood function, which cannot be computed exactly, is being approximated by "five-point adaptive Gaussian quadrature." The derivatives (gradient) of the log of this approximate likelihood are computed and the algorithm tries to find the location where the gradient vector is zero. In practice, the five-point quadrature might not be accurate enough; you can increase the number of quadrature points to get a better approximation, but the fitting procedure takes longer.

Estimation of parameters:

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper	Gradient
beta0	-5.8337	0.7473	258	-7.81	<.0001	0.05	-7.3054	-4.3621	0.000124
beta1	0.3858	0.9127	258	0.42	0.6729	0.05	-1.4115	2.1830	-0.0009
beta2	-0.3032	0.4523	258	-0.67	0.5033	0.05	-1.1939	0.5876	0.000954
beta4	0.8051	0.9392	258	0.86	0.3921	0.05	-1.0443	2.6545	0.003987
beta5	0.06371	0.9345	258	0.07	0.9457	0.05	-1.7764	1.9038	0.000246
beta6	0.1823	0.9569	258	0.19	0.8490	0.05	-1.7020	2.0667	0.004552
beta7	0.5395	0.9770	258	0.55	0.5813	0.05	-1.3844	2.4635	-0.0039
beta8	1.1784	0.9956	258	1.18	0.2376	0.05	-0.7821	3.1390	-0.00558
beta9	2.0257	0.6685	258	3.03	0.0027	0.05	0.7094	3.3421	-0.00263
beta10	3.4255	0.6792	258	5.04	<.0001	0.05	2.0880	4.7630	-0.00071
beta11	4.9886	0.7141	258	6.99	<.0001	0.05	3.5825	6.3948	-0.0031
beta12	5.3066	0.7308	258	7.26	<.0001	0.05	3.8676	6.7456	0.002068
beta13	5.3100	0.7385	258	7.19	<.0001	0.05	3.8558	6.7641	0.003391
sigma _u	2.9538	0.2933	258	10.07	<.0001	0.05	2.3763	3.5313	-0.00025

Based on the estimated parameter, we can estimate the odds ratio. For example, estimate Odds ratio of treatment group vs. placebo at visit 8:

$$\omega = \frac{p_{tr} / (1 - p_{tr})}{p_{pl} / (1 - p_{pl})} = \exp \left(\log \left(\frac{p_{tr}}{1 - p_{tr}} \right) - \log \left(\frac{p_{pl}}{1 - p_{pl}} \right) \right) = \exp (\beta_1 + \beta_8)$$

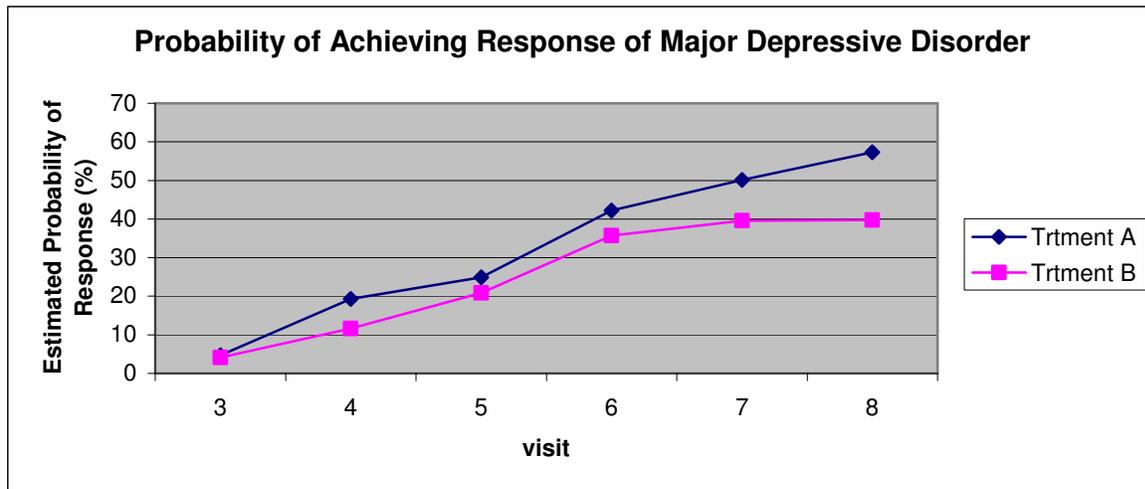
The interested population mean can be derived from the model output by integrating out the random effects. Usually this needs to be done numerically or based on sampling methods.

1. Simulate the vector of random effects $u^{l,m}$ from $N(0, \sigma^{2(m)})$ for $l = 1, \dots, L = 10000$;
2. Calculate the subject-specific probability of having a response through

$$E(Y_{ij} | u_i) = \frac{\exp(x'_{ij} \beta + u_i)}{1 + \exp(x'_{ij} \beta + u_i)}$$

3. Calculate the population-average probability of having a response by averaging each subject-specific probability.

Estimated Response Rates across Visits:



THE %GLIMMIX MACRO

The %GLIMMIX macro uses the restricted/residual pseudo-likelihood (REPL) or the pseudo-likelihood (PL) algorithms to obtain REML-like or ML-like parameter estimates of the dispersion parameter. (Wolfinger and O'Connell 1993). The macro calls the MIXED procedure iteratively until convergence, which is determined by using the relative deviation of the variance/covariance parameter estimates. An extra-dispersion scale parameter is estimated by default.

The following SAS code using the %GLIMMIX macro in the context of logistic random-effects model:

```
%glimmix (data=MMRMINV,
  stmts = %str(
    class patient poolinv visit therapy;
    model res = basval therapy visit therapy*visit /
      ddfm=kenwardroger;
    repeated visit / subject=patient(poolinv) type=un;
    lsmeans therapy*visit / cl diffs;
  ),
  error=binomial,
  link=logit,
  maxit=99
)
run;
```

All commands within the parentheses specify the various characteristics of the procedure. These include the usual commands used in PROC MIXED. The ERROR, LINK statements specify the error distribution, the link function, respectively. If you specify METHOD=REML, you get REML variance component estimates. REML estimates produce REPL estimation of the generalized linear mixed model. METHOD=ML obtains ML variance estimates of the model. The least squares mean is expressed in terms of the link function-the logit. Mu uses inverse link to express the least squares mean in terms of the original scale. The standard errors are computed on the logit scale. The output does not give the converted standard error. You can get an approximate standard error using the Delta method, which involves a Taylor series approximation. See Bishop, Fienberg, and Holland (1975) for details.

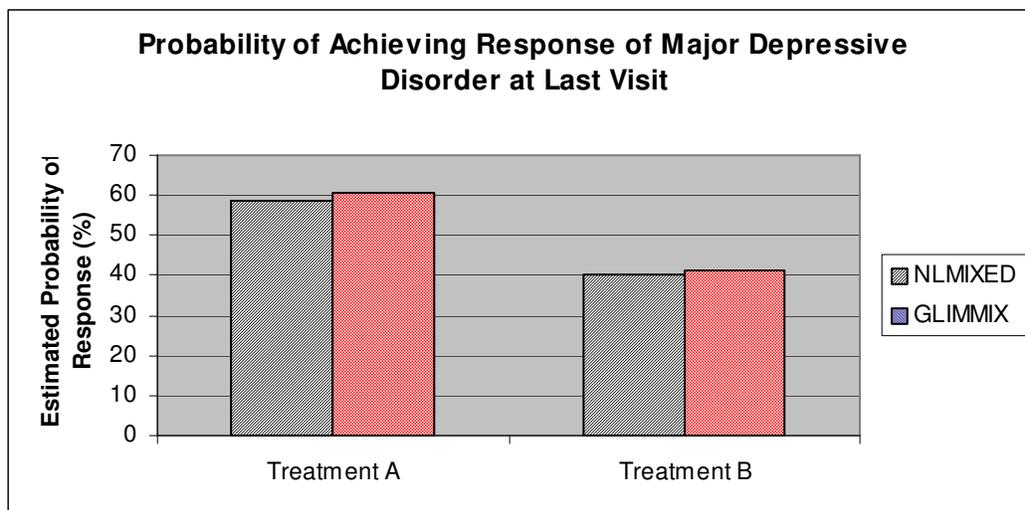
HAMD17 Total Score Response Rate
 Categorical Repeated Measures Analysis
 Analysis Based on Least Square Mean (LSM) for Within-Treatments

Therapy	VISIT	MU	CI Bound		Estimate	T	DF	p-Val Within Treatments
			Lowermu	Uppermu				
Treatment A	3	0.049	0.022	0.105	-2.967	-7.080	1280	<.001
Treatment B		0.044	0.020	0.095	-3.077	-7.350	1280	<.001
Treatment A	4	0.207	0.141	0.293	-1.345	-5.707	1280	<.001
Treatment B		0.130	0.082	0.200	-1.898	-7.256	1280	<.001
Treatment A	5	0.269	0.193	0.360	-1.002	-4.619	1280	<.001
Treatment B		0.236	0.169	0.319	-1.176	-5.533	1280	<.001
Treatment A	6	0.458	0.363	0.556	-0.170	-0.847	1280	.397
Treatment B		0.387	0.302	0.480	-0.458	-2.375	1280	.018
Treatment A	7	0.533	0.432	0.631	0.132	0.637	1280	.524
Treatment B		0.418	0.327	0.515	-0.331	-1.662	1280	.097
Treatment A	8	0.608	0.502	0.705	0.439	1.999	1280	.046
Treatment B		0.413	0.320	0.513	-0.351	-1.704	1280	.089

Model: basval therapy visit therapy*visit
 Cov.Structure=UN

CONCLUSION

In this paper, we present and describe using SAS Proc NLMIXED and %GLIMMIX macro to fit the logistic random-effects model for longitudinal binary data. The logistic random-effects model is valid only under the assumption of missingness completely at random. Both methods provide similar results.



The NLMIXED procedure lacks the CLASS statement, a variety of covariance structures that can be modeled, a repeated statement and only allows you to specify one random statement. For logistic random-effects model, the NLMIXED procedure doesn't offer greater capability of analyzed the data than %GLIMMIX. For repeated binary responses with few repeated measures on each subject, %GLIMMIX can produce biased results (Breslow and Clayton 1993).

Some key aspects of PROC NLMIXED are highlighted to demonstrate how to estimate marginal means and odds ratio of the probability of obtaining a favorable response. A SAS macro is also devised to automate fitting models. The macro allows you to write simpler syntax like Proc mixed. The syntax for the macro is the following:

```
%MACRO MMRMDB(DATA=,BVBEG=,BVEND=,CVBEG=,CVEND=,Y=,TX=);
```

```
DATA BASE COMP;
  SET &DATA;
  IF &BVBEG <= visit <=&BVEND THEN OUTPUT BASE;
  IF &CVBEG <=visit <= &CVEND THEN OUTPUT COMP;
  WHERE HAMDTL17 ne .;
run;
```

```
PROC SORT DATA=BASE;
  BY POOLINV PATIENT VISIT;
run;
```

```
data BASE ;
  set BASE ;
  BY POOLINV PATIENT VISIT;
  if Last.patient;
  Basval=HAMDTL17;
  KEEP POOLINV PATIENT INV BASVAL &TX ;
run;
```

```
PROC SORT DATA=Comp;
  BY POOLINV PATIENT VISIT;
run;
```

```
DATA MMRM ;
  MERGE BASE(in=K) COMP(in=kk);
  BY POOLINV PATIENT;
  IF K and KK ;
  IF BASVAL=0 then BASVAL=0.0001 ;
RUN;
```

```
DATA MMRM;
  SET MMRM;
  RATIO=(HAMDTL17-BASVAL)/BASVAL ;
  IF ratio =< -0.50 then &Y=1;
  IF ratio > -0.5 then &Y=0;
  IF BASVAL>=21 THEN BASCT=1;
  ELSE BASCT=0;
  RUN;
```

```
%MEND MMRMDB;
```

```
%MMRMDB(DATA=WKLY,BVBEG=1,BVEND=2,CVBEG=3,CVEND=8,Y=RES,TX=TRT);
```

```
%NLMIXLOG (DB=almost,BVBEG=1,BVEND=2,CVBEG=3,CVEND=8,Y=RES,TX=TRT,
  DUMMYVAR=BASELINE /VISIT /INVGRP,
  MODEL= BASELINE +TRT+VISIT+TRT*VISIT+ INVGRP,
  VWHERE=%STR(WHERE HAMDTL17 NE .), DELIMITER=%STR());
```

REFERENCES

Agresti, A. (2002) Categorical Data Analysis. New York: Wiley.
Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996) SAS System for Mixed Models.
Kenward, M.G., and Molenberghs, G. (1998) Likelihood based frequentist inference when data are missing at random.
Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. New York: John Wiley & Sons.
Choi, L., Dominici, F., Zeger, S.L. and Ouyang, P. (2001) Estimating Treatment Efficacy Over time: a logistic regression model for binary longitudinal outcomes. Biometrics and Reporting.
Ashby, M., Neuhaus, J., Hauck W., Baccchetti, P., Heilbrow, D., Jewell, N., Segal, M. and Fusaro, R. (1992). An annotated bibliography of methods for analyzing correlated categorical data. Statistics in Medicine 11.
SAS/STAT User's Guide version 9, SAS Institute Inc.

ACKNOWLEDGMENTS

The author thanks Shuyi Shen, Craig Mallinckrodt and Ilya Lipkovich for helpful assistance.

CONTACT INFORMATION

Your comments and questions are valued and encouraged, contact the author at:

Bin Yang
Eli Lilly and Company
1400 W Raymond Street, Bldg 140, DC 4103
Indianapolis, Indiana
Work Phone: 317-433-3114
Email: YANG_BIN_ZX@LILLY.COM

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.