How to use SAS^{Ò} for Logistic Regression with Correlated Data

Oliver Kuss
Institute of Medical Epidemiology, Biostatistics, and Informatics
Medical Faculty, University of Halle-Wittenberg, Halle/Saale, Germany

SUGI 27

SAS® Users Group International
Orlando, Florida • April 14–17, 2002

Contents

1. Introduction

2. The Data

3. The Logistic Regression Model with Correlated Data

4. Methods of Estimation in SASÒ

5. Comparison of Methods

6. Conclusion

7. References

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.

1. Introduction

Logistic regression is the standard analyzing tool for binary responses

Reasons:

- Ease of interpretation of parameters
- Prognoses for the event of interest are possible
- Software is available (LOGISTIC, GENMOD, PROBIT, CATMOD, ...)

Crucial assumption in standard logistic regression:

Observations are independent

However, many study designs in applied sciences give rise to correlated data/responses:

For example:

- Subjects are followed over time and responses are assessed at different time points
- Subjects are treated under different experimental conditions
- Several responses are measured at the same subject
- Subjects are observed in logical units (families, communities, clinics)

Analysis for discrete responses is more complicated than for continuous responses

2. The Data

Multicenter randomized controlled clinical trial, conducted in eight different clinics (Beitler/Landis, 1985, Wolfinger, 1999)

Purpose of study: Assess the effect of a topical cream treatment on curing nonspecific infections.

In each of the eight clinics, the number of treated and the number of successfully cured persons were recorded for treatment and control:

```
data infection;
    input clinic treatment x n;
    datalines;
    1 1 11 36
    1 0 10 37
        ...
    8 1  4  6
    8 0  6  7
run;
```

A crude analysis:

Ignore the fact that the data were observed in different clinics, collapse the data in a single 2x2-table, and measure the treatment effect by the odds ratio:

```
data infection2(drop=i);
     set infection;
     do i=1 to n;
        if i<= x then cure=1;
        if i > x then cure=0;
        status=2-cure;
        output;
     end;
run;

proc freq data=infection2;
          tables treatment*cure / relrisk;
run;
```

(Partial) PROC FREQ Output:

```
            Statistics for Table of treatment by cure

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
fffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
Case-Control (Odds Ratio)      1.4979        0.9151        2.4518
```

Moderate, non-significant (p=0.11) benefit for treatment. The odds for curing the infection is 50% higher in the treatment group.

However, this ignores the effect of clinics completely.

We might suspect that different features of the clinics (personnel, environment, typical population) might influence the treatment effect.

This also implies a correlation of patients from the same clinic.

3. The Logistic Regression Model with Correlated Data

There are two different groups of statistical models for binary responses that account for correlation in a different style and whose estimated parameters have different interpretations (Diggle/Liang/Zeger, 1994):

Marginal Models and **Random Effects Models**

Some Notation:

Let $\mathbf{Y_{ij}}$, (i=1,..., n, j=1,..., n_i) denote whether patient j in clinic i was cured

(Y_{ij} = 1: yes, Y_{ij} = 0: no) and

$\mathbf{x_{ij}}$ whether patient j in clinic i was in the treatment or in the control group

(x_{ij} = 1: treatment, x_{ij} = 0: control)

The response Y_{ij} is assumed to follow a Bernoulli distribution with cure probability p_{ij}.

Marginal Model

In a marginal model the treatment effect is modelled separately from the within-clinic correlation

Model equations:

$$\mathrm{logit}(p_{ij}) = b_0 + b_{treat}\, x_{ij}$$

$$\mathrm{Var}(Y_{ij}) = p_{ij}\,(1- p_{ij})$$

$$\mathrm{Corr}(Y_{ij},Y_{ik}) = \alpha$$

- Interpretation of parameters is analogous to standard logistic regression
- Correlation is regarded a nuisance parameter and is not estimated
- Correlation is assumed to be constant between patients from the same clinic and identical within clinics
- The interpretation does not depend on the single clinic but rather averages the treatment effect across clinics (? *Population-averaged*)

Random effects Model

In a random effects model it is assumed that there is natural heterogeneity between clinics and that this heterogeneity can be modelled by a probabilistic distribution

Model equation:

$$\mathrm{logit}(p_{ij} | u_i) = b_0 + b_{treat}\, x_{ij} + u_i$$

with $u_i \sim N(0,?^2)$

- To be more specific, this is actually a random intercept logistic regression model
- The correlation of patients from the same clinic arises from their sharing specific but unobserved properties of the respective clinic. *Given u_i*, the responses from the same clinic are independent.
- By considering the intercept random but the treatment effect fixed, we assume that the treatment effect is identical across clinics, but there is a single individual baseline cure probability in each clinic (? *Subject-specific*)

4. Methods of Estimation in SAS$^{\text{Ò}}$

4.1 The GENMOD Procedure

- The GENMOD procedure fits Generalized Linear Models (McCullagh/Nelder, 1989)
- Since Version 6.12 it also allows the modelling of correlated data via the REPEATED-Statement
- The implemented estimation procedure is GEE (Liang/Zeger, 1986)
- It estimates a marginal model

```
proc genmod data=infection2 descending order=data;
  class treatment clinic;
  model cure=treatment / d=bin link=logit;
  repeated subject=clinic / type=cs;
  estimate "treatment" treatment 1 -1 / exp;
run;
```

4.2 The %GLIMMIX Macro

- The %GLIMMIX macro was written by Russ Wolfinger from SAS® Institute and is available from the SAS® homepage
- It is designed for the analysis of Generalized Linear Mixed Models (GLMM)
- Our random intercept logistic regression model is a GLMM
- Several estimation methods are possible
- Iteratively fits a linear mixed model to a pseudo response (Wolfinger/O'Connell, 1993)

```
%include "...\glmm800.sas";
%glimmix(data=infection2,
  stmts = %str(class clinic;
               model cure = treatment / solution cl;
               random clinic;
               parms (0) (1.0);),
    error=binomial,link=logit,procopt=order=data
         );run;
```

4.3 The %NLINMIX Macro

- The %NLINMIX macro was also written by Russ Wolfinger from SAS® Institute and is available from the SAS® homepage
- It is actually designed for the analysis of nonlinear mixed models but as our model is also a nonlinear model, we can use it (Wolfinger/Lin, 1997)
- There were substantial changes between Version 6.12 and Version 8
- Several estimation methods are possible

```
%include "...\nlmm800.sas";
%nlinmix(data=infection2,
  model =%str(
    num = exp(b0 + b_treat*treatment + u);
    den = 1 + num;
    predv = num/den;
             ),
  parms =%str(b0=-0.7142 b_treat=0.404),
  derivs=%str(
    d_b0 = num /(den*den);
    d_b_treat = treatment*num /(den*den);
    d_u  = num /(den*den);
              ),
  stmts  = %str(
    class clinic;
    model pseudo_cure= d_b0 d_b_treat / noint solution cl;
    random d_u / subject=clinic cl solution;
                ),
  procopt=empirical,
  expand=eblup
         );
run;
```

4.4 The NLMIXED Procedure

- The %GLIMMIX and the %NLINMIX use approximations to the likehood function and thus yield only approximate ML estimators
- The NLMIXED procedure maximizes the likelihood directly by numerical integration methods (Gaussian Quadrature) and thus gives „exact“ ML estiamtors
- We also use the fact here that our model is a nonlinear mixed effect model

```
proc nlmixed data=infection2;
  parms b0=-0.7142 b_treat=0.404 s2u=2;
  eta = b0 + b_treat*treatment + u;
  expeta = exp(eta);
  p = expeta/(1+expeta);
  model cure ~ binary(p);
  random u ~ normal(0,s2u) subject=clinic;
run;
```

4.5 The PHREG/LOGISTIC Procedure

- We can also use conditional ML estimation for a random effects model
- This removes the random effect completely from the likelihood function
- It turns out that the conditional likelihood function in our case is equivalent to that one in stratified (or matched) case-control studies and so we can use the PHREG procedure

```
proc phreg data=infection2;
  model status*cure(0)=treatment /
        ties=discrete;
  strata clinic;
run;
```

However, the PHREG procedure yields only asymptotic conditional ML estimators and we can use the LOGISTIC procedure for an exact conditional analysis (Derr, 2000)

```
proc logistic data=infection2 descending exactonly;
  class clinic / param=ref;
  model cure=clinic treatment;
  exact treatment / estimate=both;
run;
```

4.6 Other methods

There are still some other estimation methods that could be used:

- Meta-Analysis (MIXED procedure)
- Nonparametric ML analysis
- MCMC (Stochastic integration, Implemented very poorly in SAS®)

5. Comparison of Methods

Back to the data set:

Method	**OR**	**[95%-CI]**
PROC FREQ	1.498	[0.915; 2.452]
PROC GENMOD	1.740	[1.102; 2.747]
%GLIMMIX	2.069	[1.176; 3.641]
%NLINMIX	2.021	[1.076; 3.794]
PROC NLMIXED	2.093	[1.162; 3.771]
PROC PHREG	2.130	[1.177; 3.855]
PROC LOGISTIC	2.130	[1.137; 4.079]

6. Conclusion

- The SAS® System offers a large number of options for estimating logistic regression models with correlated data
- It is difficult to give general recommendations which of the methods to use because this depends on (a) the data at hand and (b) on the desired interpretation of parameters (population-averaged vs. subject-specific)
- In our data set we feel most comfortable with the results from the NLMIXED procedure and from the conditional ML analysis.

7. References

- Beitler, P.J., Landis, J.R. (1985), “A Mixed-effects Model for Categorical Data,“ *Biometrics*, 41, 991-1000.
- Derr, R.E. (2000), “Performing Exact Logistic Regression with the SAS® System,“ *Proceedings of the 25th Annual SAS â Users Group International Conference (SUGI 25)*, 254-25.
- Diggle, P.J., Liang, K.-Y., Zeger, S.L. (1994), *Analysis of Longitudinal Data*, Oxford University Press, Oxford.
- Liang, K.-Y., Zeger, S.L. (1986), Longitudinal Data Analysis Using Generalized Linear Models, *Biometrika*, 73, 13-22.
- McCullagh, P., Nelder J.A. (1989), *Generalized Linear Models*, Chapman and Hall, New York.
- Wolfinger, R.D., O’Connell, M. (1993), “Generalized linear models: a pseudo-likelihood approch,” *Journal of Statistical Computation and Simulation*, 48, 233-243.
- Wolfinger, R.D., Lin, X. (1997), “Two Taylor-series approximation methods for nonlinear mixed models,” *Computational Statistics & Data Analysis*, 25, 465-490.
- Wolfinger, R.D. (1999), “Fitting Nonlinear Mixed Models with the new NLMIXED Procedure,” *Proceedings of the 24th Annual SAS â Users Group International Conference (SUGI 24)*, 287-24.