Paper 226-2008

Old versus New: A Comparison of PROC LOGISTIC and PROC GLIMMIX

Rebecca Christofferson
Department of Pathobiological Sciences LSU School of Veterinary Medicine, Baton Rouge, LA

ABSTRACT

In the past, the SAS programming tools available for logistic regression problems have been trapped in a "fixed effects" modeling world. PROC LOGISTIC gives very few options when dealing with random effects, which has made the modeling of binary data from any kind of experimental design challenging at best. Such design elements as blocking or repeated measures are not readily analyzed using PROC LOGISTIC. PROC GLIMMIX has sought to fill in the gaps. With this new procedure, design elements can be accounted for and a more correct modeling of variances can be done. The popular and useful mixed modeling techniques available in PROC MIXED can now be readily utilized for the analysis of binary data using PROC GLIMMIX. A practical example will demonstrate the convenience features GLIMMIX offers, and also highlight the differences between the two procedures, using a side-by-side comparison. Data from an experiment involving sperm morphology in game deer will be utilized for this simple demonstration.

INTRODUCTION

Control of variation is commonly the goal of statistical modeling. In the past, logistic regression using the statistical procedures available in SAS has limited statisticians and programmers to corralling logistic and binary problems into "fixed effects" models. However, modeling of random effects is important for accuracy. In the past, there was no easy way to control for variation caused by such things as experimental design or blocking effects. Recently, however, a new procedure has changed the way SAS/STAT users program for logistic regression. This paper seeks to contrast the old way (PROC LOGISTIC) with a new procedure (PROC GLIMMIX) to illustrate the differences and improvements afforded by GLIMMIX, but by no means is intended to be a comprehensive report on either procedure. Rather, this is an introduction to the exciting potential of PROC GLIMMIX and a personal show of enthusiasm for the new procedure.

What is the big deal with random effects? I've been asked this question in consulting settings, so here goes an explanation. Picture yourself sitting in a quaint, crowded sidewalk café in the dream destination city of your choice. Across from you is your vacation partner and the two of you are locked in a potentially riveting conversation (likely about SAS programming). The only problem is that you can't quite understand what your partner is saying because there is so much crowd noise. There may be multiple sources of noise, such as the other people talking, maybe a construction site not very far away, and traffic. These other sources of noise are random effects; the conversation with your partner is the model you're interested in. In accounting for random effects, you filter out that crowd noise so that all you are left with is the clearest possible model (conversation). Your ears are suddenly very discerning and you are able to easily tell the difference between what is random noise and what is conversation.

Binary data is often analyzed using logistic regression. This paper will assume that the reader has at least a basic understanding of logistic regression. By using a simple example, this paper will highlight some of the new features PROC GLIMMIX offers for modeling of binary data in comparison to those options currently available in PROC LOGISTIC.

DATA & METHODOLOGY

PROC LOGISTIC is a suitable procedure to utilize when true regression models are warranted, i.e. there are no random effects and your model is simple. However, in most research settings, experimental design is not only utilized, but it is a standard procedure. For this reason, modeling of binary data without accounting for the variation random effects may contribute has profound implications for the precision of models and predictions.

The Experiment

The data is from an experiment that sought to determine the best treatment methods for the preservation of deer sperm function via cryogenesis. The treatment in this experiment was a four level treatment; levels were pre-freeze A & B versus post-thaw A & B where A represents room temperature processing and B represents processing at a cooler temperature. They were coded as PFA, PFB, PTA, PTB and so were not done in a factorial design. Specifically, tail and head morphology was observed to determine sperm viability and thus, degree of preservation.

Several morphological features of damaged sperm were measured. If a sperm had no defects, it was assigned a "1," and if the sperm was deemed damaged, it was assigned a "0." All counts became a ratio of "success" (no damage, variable name: “normal_”) over total.

In this experiment, sperm was collected from 7 deer which were treated as a random block effect. The model was established as:

$$Y_{ij} = \mu + \beta_i + \delta_j + \varepsilon_{ij}$$

Where μ is the overall mean, β_i is the fixed effect of trt at the i^{th} level, δ_j is the random effect of the j^{th} deer, and ε is the residual error. Using PROC LOGISTIC, the basic SAS code could look as follows:

```
proc logistic data=data;
class trt;
model normal_/total = trt;
output out=pred_log;
run;
```

As is evident, there is no place in this procedure for easy address of the random effect of deer; there is no random statement available. That means that the variation for that effect is essentially unaccounted for and the model is perhaps not as precise as it could be. All that variation is pooled into the residual error, making the test for the differences between the treatment groups not as precise.

PROC GLIMMIX offers researchers the option of implementing a linear mixed model by including deer as a random effect.

```
proc glimmix data=morph12o;
class trt;
model normal_/total = trt / dist=bin solution;
output out=morphpred pred(ilink noblup)=pred resid=r ucl(ilink
                          noblup)= up  lcl (ilink noblup)=low;
lsmeans trt /adjust=tukey;
random deer / solution;
run;
```

Our model is now fully specified in the SAS coding. The **solution** option in the random statement of PROC GLIMMIX offers us a test for significance for the random effect of deer. The output of that test is show below in table 1.

Solution for Random Effects					
Effect	**Estimate**	**Std Err Pred**	**DF**	**t Value**	**Pr > |t|**
Deer	0.1115	0.005408	50	20.62	<.0001

Table 1

As you can see, the random effect of deer is significant. Again, by accounting for the variance, our ability to correctly detect differences based on this model is more precise.

The Output

The output for these two procedures looks different. However, it is important to note that the effect of *trt* is still analyzed by a difference in log-odds ratios, like PROC LOGISTIC. The parameterization is also the same, as is shown in tables 2 and 3 below.

Type III Coefficients for trt				
Effect	**trt**	**Row1**	**Row2**	**Row3**
Intercept				
trt	PFA	1		
trt	PFB		1	
trt	PTA			1
trt	PTB	-1	-1	-1

Table 2: Coefficients from PROC GLIMMIX

Class Level Information				
Class	**Value**	**Design Variables**		
trt	**PFA**	1	0	0
	PFB	0	1	0
	PTA	0	0	1
	PTB	-1	-1	-1

Table 3: Coefficients for PROC LOGISTIC using the default effect Parameterization.

As a consultant in a university setting, my clients are often graduate students who need assistance in analyzing their thesis, dissertation or other project data. Often these students have a working grasp of statistics, but when it comes to reading output, they are "stuck" on ANOVA tables. The first words out of my clients' mouths usually are, "I need to do an ANOVA" regardless of whether that is the most appropriate analysis, because everyone likes a clear, concise ANOVA table.

The output generated from PROC LOGISTIC is shown below in tables 4 & 5.

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
trt	3	199.3996	<.0001

Table 4

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
trt PFA vs PTB	2.168	1.930	2.435
trt PFB vs PTB	1.815	1.624	2.028
trt PTA vs PTB	1.480	1.327	1.650

Table 5

The type 3 Analysis of Effects is the Wald's test for the global null hypothesis β=0. This is essentially testing for a difference between the 4 groups of *trt*. The point estimates are only given as odds ratios (table 5), which is the probability of an event occurring in one group over the probability of an event occurring in another group. As I've alluded to, this output is sometimes difficult to interpret for those with only a basic level of instruction in statistics. (There is another table outputted in PROC LOGISTIC which offers maximum likelihood estimates used to calculate the estimated logit for each effect group, but these are not readily interpretable and not included here.)

In contrast, the output for PROC GLIMMIX is given below in tables 6 & 7.

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Trt	1	52	133.54	<.0001

Table 6

Notice, the type III test of fixed effects now involve the F-distribution and a p-value based on the F-distribution (table 6) which is something that most students are taught in their basic statistics courses. Explaining a new and often foreign looking set of output and tests is time consuming and frequently frustrating for the student/client. While the output should not be a huge hurtle, the reality is that GLIMMIX has made life that much easier with this familiar-looking significance test for fixed effects.

Obs	trt	Pred_Log	Low_Log	Up_Log	Pred_GLIMMIX	Low_GLIMMIX	Up_GLIMMIX
1	PFA	0.55995	0.54347	0.57629	0.56207	0.53222	0.59148
2	PFB	0.53984	0.52374	0.55585	0.52836	0.50022	0.55631
3	PTA	0.49626	0.48014	0.51239	0.47518	0.44750	0.50301
4	PTB	0.40476	0.38913	0.42059	0.37530	0.34964	0.40169

Table 7

Table 7 shows the predicted values as well as the lower and upper confidence limits from each procedure. As you can see the predicted values are very similar, but the confidence interval for PROC GLIMMIX is wider. Below is a graph depicting the confidence intervals from each of the procedures (Figure 1).

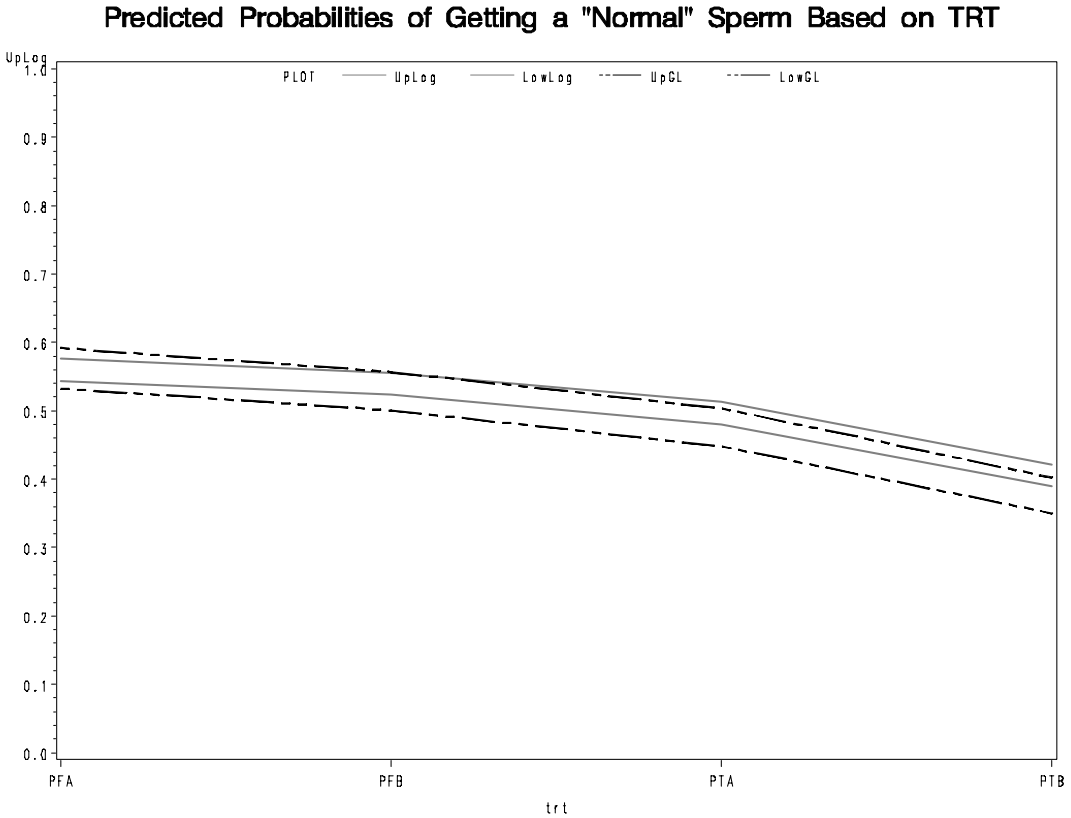

Figure 1

Again, it appears, at first glance, that PROC LOGISTIC gives a more precise prediction because the confidence interval is thinner. You must remember that when a random effect is accounted for, the variance is taken from the residual error term and "put in" the standard error of the mean (SEM) for the effect. Therefore, the SEM gets larger and confidence intervals widen. The test for the difference between groups, however, becomes better because the standard error of the difference is more accurate. Therefore your test for significance is actually improved for having a completely specified model, including random effects.

CONCLUSIONS AND PRACTICAL IMPLICATIONS

Other SAS Considerations & Convenient Options

One cannot talk about mixed modeling in SAS without mentioning PROC MIXED. In mixed modeling, PROC MIXED has become widely used. The options available, the robustness of predictions, its ability to handle missing observations, and simply the familiarity that comes with repeated use makes it a forerunner when deciding how to program for mixed models. Now, PROC GLIMMIX offers these same benefits as MIXED, but with several new and exciting options. As seen above in PROC GLIMMIX, lsmeans statements, contrast statements (available in both MIXED and LOGISTIC), even the random statement are all artifacts of the MIXED procedure. Indeed, the ease of transition to GLIMMIX from MIXED will no doubt add to its growing popularity.

There are differences that programmers should be aware of, however. The optimization method in GLIMMIX is different from that in MIXED. GLIMMIX offers many different options for optimization than does PROC MIXED, which has only one option: ridge-stabilized Newton-Raphson algorithm. GLIMMIX, by default, uses a Quasi-Newton algorithm. In addition, the MIXED procedure offers more options for specifying covariance structures than GLIMMIX currently does. The specifics of these algorithms and covariance structures are beyond the scope of this paper.

New options define GLIMMIX as the next generation of SAS programming tool for mixed models and logistic regression. One such option- and probably the most influential- is the ability to specify a distribution for the data. This option, of course, is the driving force of this example. Gone are the days when mixed modeling was only easily done with normal data. For this data, the **dist=bin** option was invoked, but almost any distribution from the exponential family can be used. A list is available in the GLIMMIX documentation.

Another feature of GLIMMIX that has made consulting life easier is the outputting of predicted values relative to LOGISTIC. In GLIMMIX, there are options. Outputting using the option **ilink** in the output statement generates the predicted values on the scale of the original data, while the default of **nolink** and is on the scale of the link function. Outputting on the scale of the original data allows for very easy use when calling data into graphical procedures, as well as for interpretation of results. In addition, in the LSMEANS statement, **ilink** will output estimates that are means (on the scale of the data), in addition to average logits.

The **noblup** option, in addition to its statistical function, in the output statement automatically outputs on a per treatment basis, and has been easily used for graphing purposes in other analyses. This is a seemingly simple thing and yet a very huge help. If using the **blup** option, each predicted observation will have its own predicted value and confidence limits. This takes into account the predicted value for the random effect, adding it into the predicted value for the mean.

Discussion

This was my first foray into the use of GLIMMIX. While I obviously have much left to learn and utilize, I feel that this was an important and exciting step. Of course, my client did not fully appreciate my enthusiasm which is why I am grateful for this type of venue. In conclusion, the marriage of these two procedures- LOGISTIC AND MIXED- has produced a new generation of statistical tool, and GLIMMIX will prove better still as it moves into standard production.

ACKNOWLEDGEMENTS
I'd like to thank Jess Saenz and Dr. Robert Godke of the Louisiana State University Reproductive Physiology lab for the use of this data. I'd also like to thank Claudia Leonardi for her editorial services, Dean Kenneth Koonce for his encouragement in using SAS and, as always, my alma mater the Louisiana State University Department of Experimental Statistics.

REFERENCES

Cook, D., Dixon, P., Duckworth, W.M., Kaiser, M.S., Koehler, K., Meeker, W.Q., Stephenson, R.W.; "Binary Response and Logistic Regression Analysis," chap.3 in *Beyond Traditional Statistical Methods*, Iowa State University, 2001.

Karp, Andrew H., "Getting Started with Proc Logisitic," in *SUGI 26 Proceedings Paper 248-26*, Long Beach, CA, April 22-25, 2001. Online.

J. R. Saenz, "Cryopreservation of White-Tail deer Epididymal Sperm for Artifical Insemination" M.S. thesis, School of Animal Sciences, Louisiana State University, Baton Rouge, LA, 2007.

Schabenberger, Oliver. "Introductin the GLIMMIX Procedure for Generalized Linear Mixed Models," in *SUGI 30 Proceedings Paper 196-30,* Philadelphia, PA, April 10-13, 2005. Online.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Becky Christofferson
Department of Pathobiological Sciences, LSU School of Veterinary Medicine
Baton Rouge, LA 70803