Paper 1275-2014

Got Randomness? SAS® for Mixed and Generalized Linear Mixed Models.

David A. Dickey, N. Carolina State U., Raleigh, NC

1. ABSTRACT

SAS® PROC GLIMMIX fits generalized linear mixed models for nonnormal data with random effects, thus combining features of both PROC GENMOD and PROC MIXED. I will review the ideas behind PROC GLIMMIX and offer examples of Poisson and binary data. PROC NLMIXED also has the capacity to fit these kinds of models. After a brief introduction to that procedure, I will show an example of a zero-inflated Poisson model, which is a model that is Poisson for counts 1,2,3,…, but has more 0s than is consistent with the Poisson. This paper was previously presented at the 2010 SAS Global Forum.

2. INTRODUCTION

This paper discusses generalized mixed models. The word "mixed" refers to models with random effects. The word "generalized" refers to nonnormal distributions for the response variable. Two of the most common distributions are the binary (Y is 1 with probability p and 0 with probability 1-p) and the Poisson. The binary distribution is useful when the outcome of an experiment or survey is a category with 2 levels such as passing or failing a test, being admitted to a hospital or not, germinating or not germinating. The statistical model parameterizes the probability p as a function of some continuous and/or classificatory explanatory variables. Because p is a probability, the function must be constrained to lie between 0 and 1. As a hypothetical example, suppose individual seeds are laid on damp soil in petri dishes and these dishes are kept at different temperatures T for various numbers of days D, a preassigned D being used for each seed. After D days, the seed is inspected and Y=1 is recorded if it has germinated, Y=0 otherwise. The probability of germinating p may depend on T and D through a linear function L=a+bT+cD where a, b, and c are parameters to be estimated. Now, unless b and c are 0 and 0<a<1, there is nothing holding L between 0 and 1. To so restrict p, it is modeled as a function of L, namely p=exp(L)/(1+exp(L)) or equivalently p=1/(1+exp(-L)). The probability that Y=0 or 1, Pr{Y=y}, is given by $p^y(1-p)^{1-y}$ which is seen to be p if y=1 and (1-p) if y=0. Multiply these p=exp(L)/(1+exp(L)) and (1-p)=1/(1+exp(L)) values together to get the likelihood function for the observed sequence of 0 and 1 responses in your data. Notice that the L values and hence the p values are different for most cases whose Ts and Ds are different. This likelihood function is complicated, but is a function of a, b, and c only so it can be maximized to give estimates of these parameters. A nice way to think of what is happening is to recall the formula L=log(p/(1-p)) that expresses L as a function of p then interpret L as an expected response and L = a + bT + cD as a model. Note that p is the expected value of Y for a binomial. The function L=log(p/(1-p)) is called a logistic link function where a link function in general links the mean of Y to a linear function L of some predictors. In Figure 1 we plot the (hypothetical) probability of a seed germinating, p, against T and D where L = -**11** + **.15***temperature + **.3***days.

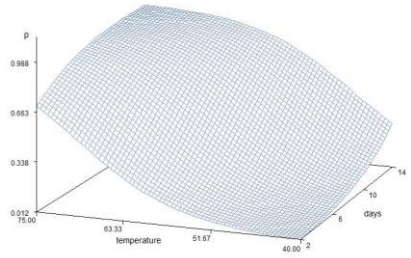

Figure 1: Germination probability versus temperature and day.

For example, if we want a probability of germination exceeding 80% we then want L to exceed log(.8/.2)=log(4) = 1.3863 so as long as **.15***temperature + **.3***days>12.2863 our model predicts this will happen, in other words we need days>41-(1/2)temp. At temperature 50 we need at least 16 days or more and at temperature 70 we expect 80% germination in 6 days. This example has no random effects so it is a generalized linear model, not a generalized mixed model.

3. LOGISTIC REGRESSION ON O-RING DATA

A real data example is provided by the US space shuttle program. Prior to the Challenger disaster, 23 missions were flown and the O-rings, the seals whose failure caused the Challenger to explode, were inspected. Each mission had 6 O-rings and after each mission, each O-ring was inspected for "erosion or blow by" where we will record a 1 if either of these forms of damage was found and 0 otherwise. To see if the 6x23=138 O-rings were affected by ambient pre-launch air temperature, we will analyze these data by logistic regression with a 0-1 response variable. Here FAILED is the number (out of 6) experiencing erosion or blowby on each mission and ATRISK is the number of O-rings (6) on each mission. These "failures" were not mission critical. These 23 shuttle missions returned safely to earth. Erosion or blowby may nevertheless signal more serious problems at more extreme levels of the predictors. A second predictor, the launch number 1 through 23 was also used to see if there was a trend in these O-ring problems over time. Using

```
PROC LOGISTIC DATA=shuttle; title3 "Logistic Regression";
 model failed/atrisk = temp launch;
 output out=out1 predicted = p; run;
```

we find that L = 4.0577 - 0.1109(temperature) + 0.0571(launch) where the p-values for the three coefficients are Pr>Chi-square = 0.1807, 0.0122, and 0.3099 so the intercept and launch number are insignificant at the 5% level. These effects are illustrated in the graph of p=exp(L)/(1+exp(L)) in Figure 2.

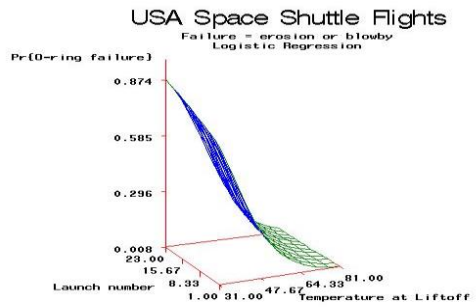

Figure 2. O-ring probability of "failure" (erosion or blowby).

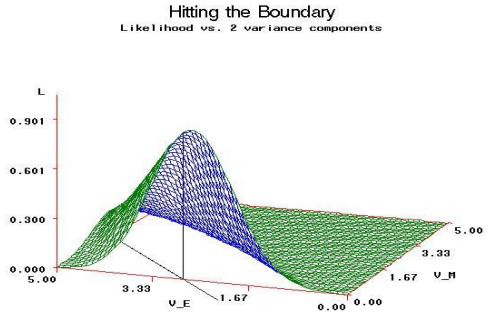

Figure 3. Boundary problems in random effects models.

The leftmost temperature coordinate in Figure 2 is 31 degrees F, the launch temperature for the last and fatal Challenger launch. The probability of erosion or blowby *on each O-ring* is $p= e^L/(1+e^L)= 0.88$ where L=4.0577-0.1109(31)+.0571(24)=1.99.

The model just given had no random effects so why is this example included in a PROC GLIMMIX presentation? Recall that these O-rings are clustered by mission. Having found no deterministic trend in these "failures," we might consider the possibility of a *random* mission effect. Some researchers describe this as clustering (missions are the clusters) in the failures. In fact we should have thought of the individual O-rings as subsamples on 23 treatment units, an experiment that has 2 variance components, one for mission and the other for individual rings within missions and a treatment of sorts, temperature. The estimation of these random variance components is done with REML in PROC MIXED, which cannot handle nonormal distributions, and with related methods in PROC GLIMMIX which *can* handle binomial data like these. This illustrates the main practical difference between using PROC MIXED and using PROC GLIMMIX. For the shuttle data we run two models and get two results, both indicating that the boundary (variance nonnegative restriction) was encountered.

```
PROC MIXED DATA=O_ring;
class mission;
model fail = launch  temp;
random mission; run;
```

```
Covariance Parameter Estimates
Cov Parm     Estimate
mission             0
Residual      0.05821
```

```
PROC GLIMMIX DATA=O_ring;
 class mission;
 model fail = launch temp
      /dist=binomial;
 random mission;  run;
```

```
Covariance Parameter Estimates
Cov                    Standard
Parm     Estimate       Error
mission    3.47E-18
```

A problem often encountered in these iterative methods is "hitting the boundary" by which I refer to a constrained (variances must be non-negative) search over a likelihood or other objective function in which one of the parameters,

a variance in our example, is driven to the boundary of its parameter space. This is illustrated in Figure 3 above where the (hypothetical, not from the data) objective function is increased by decreasing both V_M, the mission variance on the right hand axis, and V_E, the error or within mission variance component on the right-to-left axis in the floor. Notice that the objective function increases as V_M decreases ultimately to 0 where a vertical line marks the maximum of the likelihood within the constrained region. Had the boundary constraint been removed it appears that V_M would have moved to a negative value and V_E would decrease further than the value near 3 as shown by the reference line in the plot's floor. Thus hitting the boundary with one parameter can cause other parameter estimates to disagree with those from unconstrained methods like the method of moments estimators often applied to extract variance components from mean squares in early statistical texts.

With the mission variance estimate being right on the null hypothesis value 0, we feel comfortable in saying that there is no evidence of a mission effect so that the original PROC LOGISTIC approach seems sufficient. No reasonable testing procedure would reject the null hypothesis when the estimate is exactly the hypothesized value. Note also that only GLIMMIX was able to treat the response as binomial whereas MIXED incorrectly treats the 0-1 variable as normally distributed, though this is of little consequence here.

4. A BINOMIAL EXAMPLE WITH NONTRIVIAL RANDOM EFFECTS

The data in this section are from the Centers for Disease Control. During flu season, blood samples are tested for active flu virus. The proportion p of samples containing active virus is a function of the time of year. The data are weekly. For our purposes, we say that flu season starts in week 40 of the reference year then extends across year boundaries so there is some data manipulation and accounting that must be done to organize the data set. For example, our data set has both week of the year and week into flu season. Figure 4 is a plot of p on the left and the logit link L=ln(p/(1-p)) on the right, versus week into flu season with different segmented lines for each year and a vertical reference line at the last day of December, where the flu season crosses the calendar year.

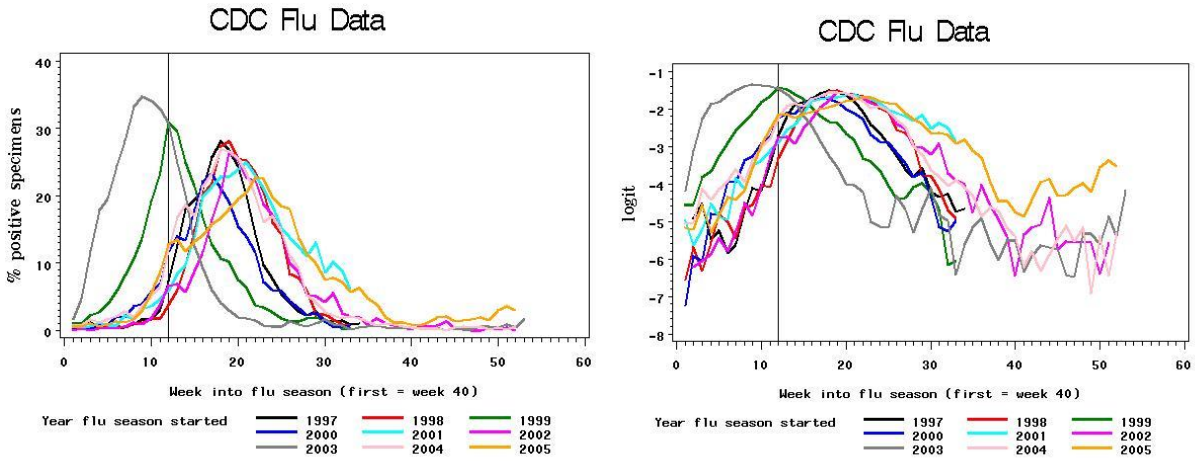

Figure 4. Flu prevalence

Note the logits on the right. There is a sinusoidal nature to the plots with the phase and possibly also the amplitude varying a bit with year. This suggests a sine and cosine wave of period 52 weeks (using both allows the coefficients to adjust the amplitude and phase to fit the data). Also note that the data do not always cover 52 weeks. In several of the years, data collection stopped at about 35 weeks after flu season started, this being essentially the end of that flu season. Harmonics (sine and cosine pairs that go through multiple cycles per year) can be introduced to modify the shape or "wave form" of the predicted values. Some wiggling in the forecasts for years in which no data were available after week 35 will result from the inclusion of these. We intend to use year as a variable but care must be taken to use the year that the flu season started, not calendar year. Otherwise unwanted discontinuities across calendar years would occur. Four analyses will be presented

4.1 FIXED EFFECTS

Here we use sines and cosines, S1 and C1 at the fundamental frequency of one cycle per year and S2, C2, S3, C3 etc. as sine cosine pairs going through 2, 3, etc. cycles per year. Here is the GLM code and resulting graphs, Figure 5.

```
PROC GLM DATA=FLU;  class fluseasn;
model logit = s1 c1
fluseasn*s1 fluseasn*c1 fluseasn*s2 fluseasn*c2
fluseasn*s3 fluseasn*c3 fluseasn*s4 fluseasn*c4;
output out=out1 p=p;
data out1; set out1; P_hat = exp(p)/(1+exp(p));
label P_hat = "Pr{pos. sample} (est.)";   run;
```

One interesting feature here is that only the harmonics are allowed to change with fluseasn, the year the flu season started. Adding main effects of the harmonics into the model did not seem to help much and note that this is an exception to the usual rule of including any main effects that are part of an interaction. The fundamental wave is the same across years with variations in the wave form changing year to year. Figures 5 and 6 are plotted near each other here for easy visual comparison between this fixed approach and the upcoming random effect approach.

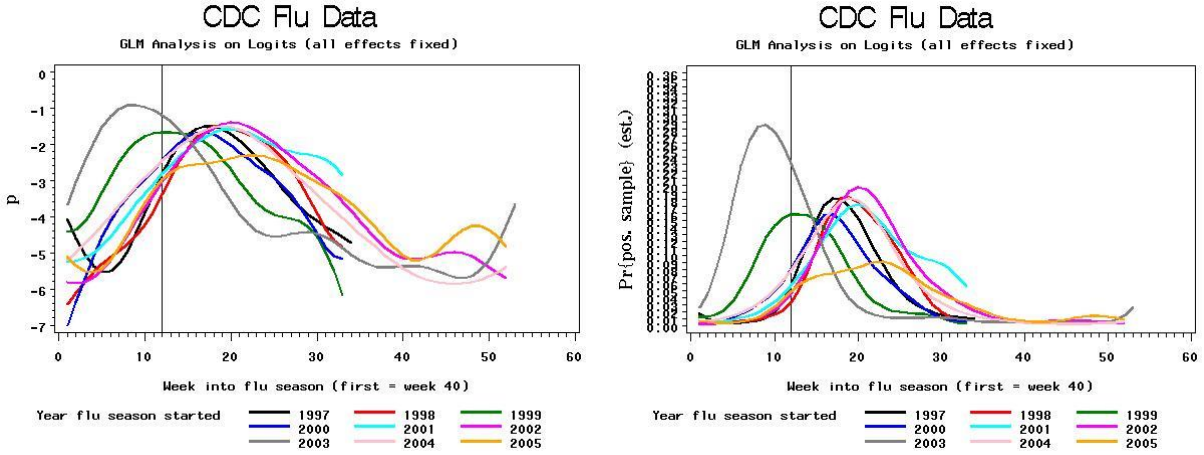

Figure 5. Predicted logits and proportions, PROC GLM.

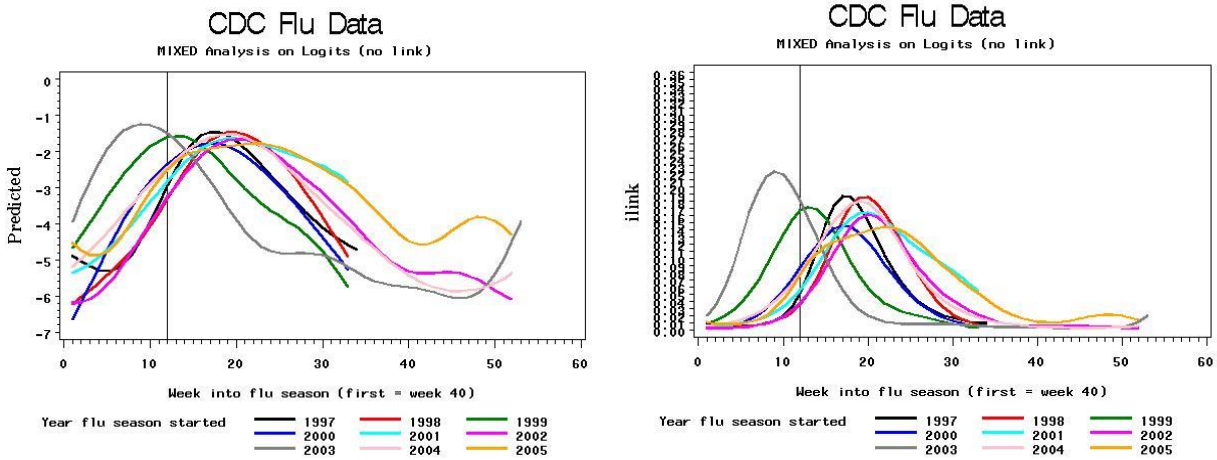

Figure 6. Predicted logits and proportions, PROC MIXED.

4.2 MIXED MODEL ANALYSIS

The model with years considered random, and run in PROC MIXED, gives Figure 6. Figures 5 and 6 are plotted near each other above for easy visual comparison between the fixed and random effect approaches. As you compare Figures 5 and 6, notice how the assumption that the random effects are all drawn from the same distribution across years modifies the more extreme curves in Figure 5 to more similar curves in Figure 6. The unusually high peak in 2003 is lowered and the low peak in 2005 is raised. This is typical of the best linear unbiased predictors (BLUPs) produced in mixed model analysis. The code appears here:

```
PROC MIXED DATA=FLU method=ml;
 class fluseasn;
 model logit = s1 c1 /outp=outp outpm=outpm ddfm=kr;
 random intercept/subject=fluseasn;
 random s1 c1/subject=fluseasn type=toep(1);
 random s2 c2/subject=fluseasn type=toep(1);
 random s3 c3/subject=fluseasn type=toep(1);
 random s4 c4/subject=fluseasn type=toep(1); run;
```

The type=toep(1) command implies a 2x2 toeplitz matrix with only one nonzero entry in the first row, that is, we have an iid assumption on each sine cosine pair. Notice that neither the GLM nor the MIXED approach enforces a nonnormal distribution on the data nor do they enforce the relationship between the mean and the variance that often occurs in nonnormal distributions. To do this, we will need PROC GLIMMIX.

4.3 A GENERALIZED LINEAR MIXED MODEL FOR FLU PREVALENCE

In order to have a generalized model, some decision must be made about the type of distribution to be used. The data are proportions of blood samples showing viral activity. Each of these could be considered a binomial sample from a set of n trials where n is the number of blood samples tested. The data set contains the actual counts of blood samples tested (SPECIMENS) and the number testing positive for virus (POS) so the POS/SPECIMENS notation can be used to automatically inform PROC GLIMMIX that these data are binomial and thus that the logit link should be used unless the user intervenes. Here is the code:

```
PROC GLIMMIX DATA=FLU; title2 "GLIMMIX Analysis";
class fluseasn;
model pos/specimens = s1 c1 ;
random intercept/subject=fluseasn;
random s1 c1/subject=fluseasn type=toep(1);
random s2 c2/subject=fluseasn type=toep(1);
random s3 c3/subject=fluseasn type=toep(1);
random s4 c4/subject=fluseasn type=toep(1);
random _residual_;
output out=out2 pred(ilink blup)=pblup pred(ilink noblup)=overall
pearson=p_resid; run;
class fluseasn;
```

Some preprocessing, in which the fixed effects of the harmonics were included, led to the conclusion that the fundamental frequency using S1 and C1 was sufficient for the fixed effects in the model. The idea of assuming the sine and cosine components of each pair have the same variance, as enforced by the toep(1) covariance structure, comes from the study of spectral analysis in time series. It has a nice theoretical justification but would not necessarily hold when real data are analyzed nor would its failure to hold cast suspicion on the analysis. It is simply a justification for some reduction in the number of fitted parameters.

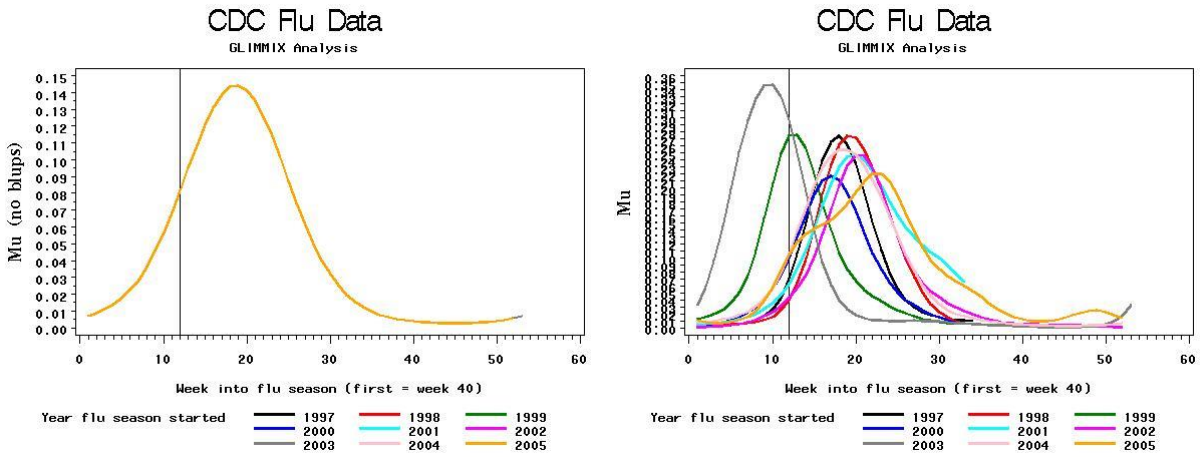

Figure 7. PROC GLIMMIX analysis illustrating the effect of BLUPs.

The graphs shown in Figure 7 are both on the scale of p, the inverse link or "ilink" scale but the left one does not include the random effects and the one on the right does. On the left the pattern is the function that is in common to all years. Comparing to that on the right, we see the effect of allowing the harmonics to change in a random fashion from year to year where year here is the year the flu season began.

Note the use of the random _residual_ command. This allows for overdispersion. Analysis of the Pearson residuals showed their variance, 3.6257, to be nontrivially greater than 1, implying overdispersion. Recall that in a binomial distribution with n trials, the fraction f of successes has mean p and variance p(1-p)/n, that is, the binomial assumption forces a relationship between the mean and variance of the data that might not hold in the data being analyzed. The Pearson residuals use this relationship to produce residuals that would have variance 1 if the model and distribution were as specified. Reading about the data used here, it appears that each week's number is a composite from different labs. There is no breakdown by individual lab and thus no way to add a random lab effect into our model. It seems, at least to this author, reasonable to assume some lab to lab variation which could easily cause overdispersion. The command under discussion, random _residual, just inflates the standard errors of test statistics in a way suggested by the ratio of the variation in the data to the assumed binomial variance. Recall that some fixed effect harmonics were eliminated earlier, illustrating the importance of having correct standard errors for tests.

4.4 A BETA DISTRIBUTION MODEL

Another approach, given the overdispersion, would be to use a different distribution on the [0,1] interval, such as the beta distribution, to model the data. The beta distribution has four parameters. The interval [0,1] is, in general, [theta, theta+sigma] so we want theta=0 and sigma=1. The other two parameters, alpha and beta, will be estimated. The binomial has one parameter and thus both the mean and variance must be functions of that one parameter and hence related to each other. With 2 parameters it is not necessarily the case that the variance depends in a critical way on the mean. Here is the code for the beta analysis where F is the sample fraction F=POS/SPECIMENS .

```
PROC GLIMMIX DATA=FLU; title2 "GLIMMIX Analysis";
 class fluseasn;
 model f = s1 c1 /dist=beta link=logit s;
 random intercept/subject=fluseasn;
 random s1 c1/subject=fluseasn type=toep(1);
 random s2 c2/subject=fluseasn type=toep(1);
 random s3 c3/subject=fluseasn type=toep(1);
 random s4 c4/subject=fluseasn type=toep(1);
 output out=out3 pred(ilink blup)=pblup pred(ilink noblup)=overall
 pearson=p_residbeta; run;
```

The Pearson residuals have these variances:

Binomial, no random _repeated_ statement	3.6257
Binomial with a random _repeated_ statement	0.83150
Beta	0.82201

Both the binomial, with a random _repeated_ statement, and the beta have Pearson residual variances close to enough to 1. The decision between them should be made on other grounds. For example, the fact that the binomial adjustment is an ad hoc standard error inflation whereas the beta fits a precise distribution would favor the beta which happens to have the smallest Pearson residual variance by the slimmest of margins. The resulting graphs are omitted as they are almost identical to those in Figure 7.

Finally, notice that all of these analyses use the logistic link and then back transform. The comparison of section 4.1 to 4.2 shows the effect of considering years as random versus fixed while ignoring the binomial nature of the data. In large samples the binomial and normal distributions are quite alike. Logically, the last two analyses, binomial and beta, are appealing for their consistency with the nature of the data. Notice also that the fixed effect sinusoid when plotted is continuous. The random effect adjustments to it would not plot as continuous functions if the data were plotted over total time instead of by year. There is nothing enforcing continuity as we cross from one flu season to the next. Because there is a substantial gap between flu seasons this is not a problem here but might render this model inappropriate in other time series situations.

5. POISSON EXAMPLES

The Poisson probability function is appropriate for count data with no upper bound to the range, that is, Y could be 0,1,2, … We have the formula $\Pr\{Y=j\} = \exp(-\lambda)\lambda^j/j!$. For such distributions λ is the mean and the variance is also λ. This λ is often referred to as the Poisson "intensity" as it is the average event count. As an example we reference a data set from the course notes entitled Statistical Analysis with the GLIMMIX Procedure (2007). The data are on nesting female horseshoe crabs. Nesting females attract males in addition to their mates. These males are called "satellites". The study looks into what features of the females tend to affect the number of satellites. One feature is the female's carapace width and another is the weight of the female. The female nests were found at various sites with several nests per site. Our interest lies in modeling λ as a function of carapace width, weight, and a random site effect. Using a linear model for λ could result in negative intensities, $\lambda < 0$, which would make no sense. The natural link function for a Poisson is the logarithm, so we will model $L = \ln(\lambda)$ as a linear function of our inputs, one of which is a random variable. This combination of a link function and random effects leads us to using PROC GLIMMIX. As a first try, we fit the model just suggested.

```
PROC GLMMIX DATA=crab;
 class site;
 model satellites = weight width / dist=poi solution ddfm=kr;
 random int / subject=site;
 output out=overdisp pearson=pearson; run;
PROC MEANS DATA =overdisp n mean var;
 var pearson; run;
```

The Pearson residuals have variance 2.6737, that is, the variance is more than twice what would be expected if the mean and variance were the same value λ, as must be the case for truly Poisson data. A quick glance at the data shows that there are a lot of female nests with 0 satellites. This suggests a distribution called the Zero Inflated Poisson for which an associated model would be termed a "ZIP" model. The idea is that with probability p_0, Y is 0 and with probability $(1-p_0)$ Y is a random value from a Poisson distribution with some intensity λ. The mean of Y is then $\mu = 0(p_0) + \lambda(1-p_0) = \lambda(1-p_0)$. Now $E\{Y(Y-1)\}$ is

$$0\,p_0 + (1-p_0)\sum_{j=0}^{\infty} j(j-1)e^{-\lambda}\lambda^j / j! = (1-p_0)\lambda^2 \sum_{j=2}^{\infty} e^{-\lambda}\lambda^{j-2}/(j-2)! = (1-p_0)\lambda^2$$

because the last sum is the sum of Poisson probabilities over all possible values and thus must be 1. Since

$E\{Y(Y-1)\} = E\{Y^2\} - E\{Y\} = E\{Y^2\} - (1-p_0)\lambda$ it must be that $E\{Y^2\} = (1-p_0)\lambda^2 + (1-p_0)\lambda$ and so the variance of Y is $E\{Y^2\} - [E\{Y\}]^2 = (1-p_0)\lambda^2 + (1-p_0)\lambda - [(1-p_0)\lambda]^2 = \lambda^2(1-p_0)(1-(1-p_0)) + (1-p_0)\lambda =$ $(1-p_0)\lambda(1+\lambda p_0) = \mu(1+\lambda p_0)$ showing that the variance exceeds the mean. The distribution is overdispersed compared to the Poisson. As seen, one choice is to add a random _residual_ statement to inflate the standard errors while leaving the fixed coefficient estimates unchanged. Another would be to fit the ZIP distribution but it is not one of the standard (exponential family) distributions supported by PROC GLIMMIX, however if one can specify the likelihood with a formula, then NLMIXED can do the fitting. Here is how it's done for the crab data:

```
PROC NLMIXED DATA=crab;
 parms b0=0 bwidth=0 bweight=0 c0=-2 c1=0 s2u1=1 s2u2=1;
 x=c0+c1*width+u1; p0 = exp(x)/(1+exp(x));
 eta= b0+bwidth*width +bweight*weight +u2;
 lambda=exp(eta);
   if satellites=0 then
   loglike = log(p0 +(1-p0)*exp(-lambda));
  else loglike =
   log(1-p0)+satellites*log(lambda)-lambda-lgamma(satellites+1);
   expected=(1-p0)*lambda; id p0 expected lambda;
 model satellites~general(loglike);
 Random U1 U2~N([0,0],[s2u1,0,s2u2]) subject=site;
 predict p0+(1-p0)*exp(-lambda) out=out1; run;
```

As with PROC MIXED, initial values are given for the parameters, including the error variances s2u1 and s2u2 for the normally distributed random site effects U1 (for the probability p_0) and the random site effects U2 (for the link $\eta=\ln(\lambda)$) . Note that p_0 can depend on something, carapace width and random site effect in our case, through a logit link function. Note also the statement "if satellites=0 then loglike = log(p0 +(1-p0)*exp(-lambda)); " which takes into account that 0 values can come from the 0 inflation mechanism as well as from the Poisson distribution. Note also that lgamma(satellites+1) is the logarithm of j! for satellite count j. The results show that the logit for p_0 is estimated as 13.3739 – 0.5447(width). A prior analysis (not shown) indicated that weight was not significant for p_0. The Poisson intensity λ is estimated as exp(2.7897 -0.09442 width +0.4649 weight) where the p-values for these parameter estimates are 0.03, 0.13, and 0.04. Interestingly the Poisson intensity seems to be just a function of weight while the 0 inflation probability p_0 seems to be just a function of carapace width with narrower widths producing higher estimated excess 0s and higher body weights producing higher satellite count intensities.

Of course the ZIP distribution is not the only overdispersed distribution possible. A common correction for overdispersion in the Poisson is to switch to a negative binomial distribution and might be a logical choice under an assumption on the nature of crab mating behavior. I am grateful to C. Truxillo for suggesting the following possible motivation for the negative binomial. Suppose males approach the female nest sequentially and suppose each male has a probability p of being selected as a mate. If those who are rejected still remain near the nest as satellites then the number of satellites becomes the number of trials in a binomial sequence before the first success. This is the geometric distribution, a special case of the negative binomial which, in general, is the number of trials before the k^{th} success. Under the entertained scenario, we would expect k close to 1. In SAS PROC GLIMMIX, the scale parameter estimates 1/k where k is the target number of successes. There is also a relationship between the negative binomial and the Poisson that suggests the Poisson as a certain kind of limiting case for the negative binomial.

The code appropriate for fitting a negative binomial is as follows:

```
PROC GLIMMIX DATA=crab;
   class site;
   model satellites = weight width / dist=nb solution ddfm=kr;
   random int / subject=site;
   output out=overdisp2 pearson=pearson; run;
PROC MEANS DATA=overdisp2 mean var; var pearson; run;
```

The output below shows that weight, but not carapace width, is a predictor of satellite counts.

Fit Statistics

-2 Res Log Pseudo-Likelihood	539.06
Generalized Chi-Square	174.83
Gener. Chi-Square / DF	1.03

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	site	0.09527	0.07979
Scale		0.7659	0.1349

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-1.2022	1.6883	168.5	-0.71	0.4774
weight	0.6759	0.3239	156.6	2.09	0.0386
width	0.01905	0.08943	166.2	0.21	0.8316

The statistic "Gener. Chi-Square/DF" is similar to, but not quite the same as the test statistic typically compared to 1 to assess distributional appropriateness. It is computed, as seen, on a "pseudo likelihood" and hence is not rigorously justified as a fit measure. It is recommended in these generalized linear mixed model cases using pseudo-likelihood that instead of comparing this statistic to 1, the user should instead directly estimate the variance of the Pearson residuals and compare that to 1, hence the code shown includes statements that output these residuals and pass them into PROC MEANS. The variance is given as 0.9809 so in this particular case the change from 1.03 was not of much consequence though that may not be true in general. Our scale estimate 0.7659 seems consistent with the 1/k=1 hypothesis, as it has a standard error 0.1349 and is thus somewhat less than 2 standard errors from the hypothesized value 1.

6. POPULATION AVERAGE VERSUS INDIVIDUAL SPECIFIC MODELS

In reading about generalized mixed models, one encounters the idea of population average versus individual specific models. To try to shed some light on this concept, let us imagine a scenario in which typists type a document several times each and the (Poisson distributed) number of errors is counted for each trial. Suppose we assume random typist effects, normally distributed with some variance. How can we incorporate these into a model? One possibility is to use this random effect in the Poisson intensity parameter which then satisfies $\ln(\lambda_i)$= ln(mean of Poisson) = $\mu+U_i$ for typist i. For the Poisson, the variance and the mean are related so the intensity λ controls both and they vary from typist to typist. If a random variable Y is such that its natural logarithm X satisfies $X \sim N(0,\sigma^2)$, we say that Y has a "lognormal" distribution and, by computing the expected value of Y (e^X) as

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int e^{x} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi\sigma^2}}\int e^{-\frac{1}{2}\left(\frac{x^2-2(\mu+\sigma^2)+(\mu+\sigma^2)^2-(2\mu+2\sigma^4)}{\sigma^2}\right)} dx = e^{\mu+\sigma^2/2}\left[\frac{1}{\sqrt{2\pi\sigma^2}}\int e^{-\frac{1}{2}\left(\frac{(x-(\mu+\sigma^2))}{\sigma}\right)^2}\right] = e^{\mu+\sigma^2/2}$$

we find that the mean of Y is $e^{\mu+\sigma^2/2}$. This means the mean of Y over the population is not the same as the intensity (mean of Y) for the "typical" typist, that is, the theoretical typist whose random effect U is the population average effect U=0. That "typical" typist has mean errors E{Y} = $\lambda = e^{\mu}$. This inconsistency between the average number of errors over all typists and the mean for the "typical" typist occurs because the random effect is part of, or "inside," the link function. For this reason, models with only fixed effects inside the link, or more generally within nonlinear functions, are called population average models whereas models with random variables inside nonlinear functional forms or link functions are called individual specific models and their means over the whole population are typically not just the fixed part of the model.

We can generate data from such a model in a SAS data step. Doing so with $\mu = 1$ and $\sigma^2 = 1$ for 8 typists gives the shorter vertical line segments in Figure 8 that denote the individual typist means. These are the randomly distributed (lognormal) Poisson intensities. The larger vertical line is at λ = e = 2.7183, the typical typist intensity λ when μ=1. The distributions shown are for λ=1,2,…,7 (rather than for the randomly generated intensities λ_i). Notice how the mean, variance, *and shape* of the distributions change with changing λ. In line with theory, we see that as λ increases, the distributions get closer to normality.

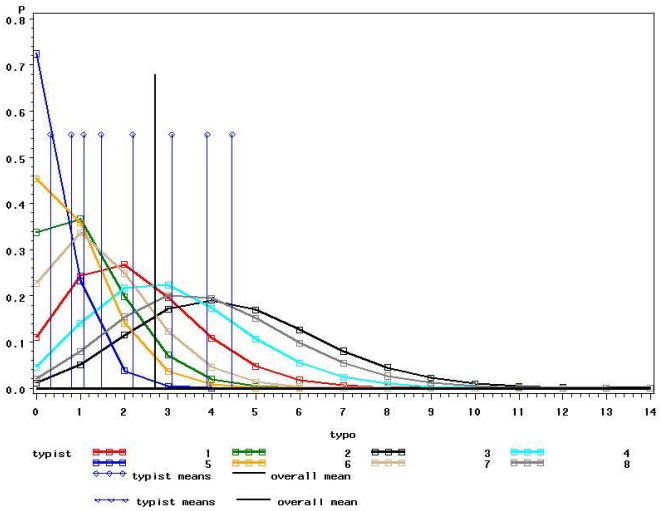

Figure 8. Typist error means and some Poisson distributions.

Note that the Poisson is a discrete distribution so that the connected dots really indicate heights of the vertical bars that would usually be used in picturing a discrete random variable's probability function. The statements made above are more clearly illuminated with the connected points.

With a program at hand that can generate intensities having random typist effects, it is now possible to increase the number of typists substantially from 8 to see what will happen. Doing this for 8000 typists and computing the average gives

The MEANS Procedure

Variable	N	Mean	Std Dev	Std Error
lambda	8000	4.4280478	6.0083831	0.067175

Why is the mean 4.4280? Recall that we have $\mu = 1$, $\sigma^2 = 1$, and thus $e^{\mu+\sigma^2/2} = e^{1.5}$ which is 4.4817, about 1 standard error away from the estimate computed in PROC MEANS above. We see that lessons learned in linear models, such as the expected value of Y being the model with the random terms omitted, do not always carry over into nonlinear (including generalized linear) mixed models, again pointing out how these models differ from population average models in which only fixed effects appear inside the link function.

7. CONCLUSION

With modern computing methods, a large variety of statistical analyses are now available that previously were unrealistic with hand computations or even with early computers. The SAS System allows implementation of thee methods in a modern computing environment. Two particular features are illustrated in this paper. Distributions

which were sometimes approximated with normal approximations in earlier times can now be estimated using generalized models with link functions. Several of these were presented herein. Even with link functions, linear models on the scale of that link are still somewhat restrictive. Models nonlinear on the link scale can be handled with SAS PROC NLMIXED and this second feature is also illustrated with an example. Distributions that are not in the exponential family, this being a family of distributions for which generalized *linear* models are well motivated, can also be fit. For example, a zero inflated Poisson model was shown.

Along with being able to handle more sophisticated models comes a responsibility on the part of the user to be informed on how to use these advanced tools. Certain things that have come to be expected in linear models or even generalized linear models may not hold in generalized linear mixed models and nonlinear mixed models. Some of the distributions discussed have variances that are functions of the mean, a property not shared by the normal distribution. In this sense the normal and some other 2 parameter distributions are more flexible for model fitting. Practitioners can encounter data sets for which a distribution at first seems reasonable but the relationship between the mean and variance does not seem to hold empirically in the data. The use of Pearson residuals for model checking was discussed. Adjustments, such as zero inflation and switching distributions were illustrated. The distinction between population average models and individual specific models, a distinction that does not arise in fixed effect models or linear mixed models, was made and illustrated with a Poisson typing error example.

REFERENCE:

Statistical Analysis with the GLIMMIX Procedure . (2007). Tiao, Jill, SAS Institute, Cary. N.C.

APPENDIX: THE REML ESTIMATION METHOD

Models often have random effects, that is, the effects are assumed to be a random selection from a normal distribution of such effects. Shuttle mission effects might be such. In estimating variances a method called REML (Residual or Restricted Maximum Likelihood) is used in PROC MIXED and a pseudo REML method called RESPL is used in PROC GLMMIX. We look first at an intuitive explanation of REML. Imagine a sample of 4 numbers {3, 7, 10, 12} with sample average 8, deviations {-5, -1, 2, 4} and sum of squares 25+1+4+16=46. The maximum likelihood method estimates the variance as (sum of squares)/n = 46/4=11.5 but this estimate is biased downward. The sum of squares for n observations is $\sum_{i=1}^{n}(Y_i-\mu)^2 - n(\bar{Y}-\mu)^2$ and dividing by n makes the first term unbiased. It is the fact that the sample mean fits the data better than the true mean μ that gives the bias $-E\{(\bar{Y}-\mu)^2 = -\sigma^2/n$. If the mean were known then dividing by n would be fine. We can always find (n-1) numbers Z with *known* mean 0 and the same sum of squares and theoretical variance as the n Y values. This motivates the division of the Z sum of squares by the number of Zs, which is n-1. To illustrate, note that vector **Y** satisfies **Y**~N(**0**,σ^2**I**). Now compute

$$TY = \begin{pmatrix} .5 & .5 & .5 & .5 \\ -.5 & -.5 & .5 & .5 \\ -.5 & .5 & -.5 & .5 \\ .5 & -.5 & -.5 & .5 \end{pmatrix}\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = \begin{pmatrix} Z_0 \\ Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} \sim N\left(\begin{pmatrix} 2\mu \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix}\right).$$

Ignoring Z_0 we have {Z_1, Z_2, Z_3} independent and identically distributed with known mean 0, and because **TT'=I**, a 4x4 identity matrix these Z's also have variance σ^2. In general, PROC MIXED assumes the model $Y = X\beta + Z\Gamma + E$. The column of residuals R in the least squares regression of Y on X is a collection of n linear combinations R = $\left(I - X(X'X)^{-1}X'\right)Y$ = $\left(I - X(X'X)^{-1}X'\right)(Z\gamma + E)$ of the vector of random terms $(Z\gamma + E)$ where X is the rank k fixed effect regression matrix and R thus has mean 0 and a known covariance matrix. Ignoring k of the residuals eliminates linear dependencies, and we have a set of variables with known mean 0 and variance matrix a known function of the variance components. Maximizing the resulting "residual likelihood" with respect to

these variance components gives the REML estimates. These are sometimes unbiased, as in our small example here, and in general are less biased than the full blown maximum likelihood estimates on the unadjusted original data. Because a true variance component could be 0, it is clear that unbiasedness cannot hold in general. For a 0 variance component the positive estimates have no corresponding negative estimates to balance them out and give average estimate 0.

An additional problem for *generalized* linear mixed models is the nonlinear link function. Thus there is no X matrix to work with and no way to compute a projection matrix $\left(I - X\left(X'X\right)^{-1}X'\right)$ that gives a known mean 0 to the transformed data. One general approach to estimating parameters of nonlinear models is to do a local Taylor Series linearization. This in effect replaces matrix X with a matrix F of partial derivatives and the likelihood of this locally linearized model approximation forms a "pseudo likelihood" to be maximized. If we choose to apply the residual likelihood method described above to this model approximation we call this residual pseudo likelihood maximization or RESPL, the default method in GLIMMIX when models with random effects are encountered. The method is iterative in that the matrix F of derivatives is a function of the current parameter estimates as we perform the search and update parameters. Because this is a pseudo likelihood, it is suggested that the fit statistic "Pearson Chi-Square/df" be replaced by the variance of the Pearson residuals when evaluating model adequacy, these values being available from the output = statement.

CONTACT INFORMATION

Name: Professor David A. Dickey
Enterprise: Department of Statistics
Address: Box 8203, North Carolina State University
City, State ZIP: Raleigh, NC 27695-8203
E-mail: dickey@stat.ncsu.edu
Web: http://www4.stat.ncsu.edu/~dickey/