Statistical Programs
College of Agricultural and Life Sciences
University of Idaho

Notes on Modeling Non-Normal Data

Terminology

LM: Linear Model. Assumes a fixed linear process fitted to Normal Data. (PROC GLM)

LMM: Linear Mixed Model. Assumes a linear process with fixed and random components fitted to Normal data. (PROC MIXED)

GLM: Generalized Linear Model. Assumes a fixed linear process fitted to Normal **or** non-normal data. (PROC GENMOD) **Note: This is different than PROC GLM!!**

GLMM: Generalized Linear Mixed Model. Assumes a linear process with fixed and random components fitted to Normal **or** non-normal data. (PROC GLIMMIX)

NLM: Non-linear Model. Assumes a fixed non-linear process fitted to Normal Data. (PROC NLIN)

GNLM: Generalized Non-linear Model. Assumes a fixed non-linear process fitted to Normal **or** non-normal data. (PROC NLMIXED)

GNLMM: Generalized Non-linear Mixed Model. Assumes a non-linear process with fixed and random components fitted to Normal **or** non-normal data. (PROC NLMIXED)

How does Normal compare to other distributions

Distributions	Mean	Variance	Nature of Data
Normal	μ	σ^2	Continuous
Binomial	Np	Np(1-p)	Count from total
Negative Binomial	μ	$\mu+\phi\mu^2$	Count
Poisson	μ	μ	Count
Exponential	μ	μ^2	Continuous
Beta	p	$p(1-p)/(1+\phi)$	Continuous in [0, 1]

For other distributions the variance is a function of the mean; The Normal is a special case. In addition, some distributions have an extra parameter, ϕ, called a scale parameter. These characteristics will influence how statistical estimation is formulated and the GLMM model will be used to carry out the estimation.

Modeling Scales

Each of these distributions has an intrinsic transformation associated with it referred to as a *link*. For example, the Poisson or Exponential distributions use the natural logarithm as their link, i.e. Link = $g(Y_{ij}) = \ln(Y_{ij})$. The Normal distribution has a special link called *identity*, which means no transformation at all, i.e. Link = $g(Y_{ij}) = Y_{ij}$. The Binomial distribution has another link called a logit and is defined as Link = logit = $g(p_{ij}) = \ln(p_{ij}/(1-p_{ij}))$, also referred to as the log odds. Link functions are used for statistical modeling and all statistical inferences are made on the linked scale, that is, the *Model Scale*. These are then translated back to the original *Data Scale* for reporting. It is important to remember that the link functions are not one-to-one transformations, and as such, the results on the Model Scale may or may not have practical applicability in the Data Scale. Hence, caution is necessary for interpreting any subsequent inferential results.

Common Statistical Models

$g(Y_{ij}) = G + \varepsilon_i$	Base Mean Model
$g(Y_{ij}) = G + T_j + \varepsilon_{ij}$	Completely Random Design (CRD)
$g(Y_{ij}) = G+ T_j + \beta_i + \varepsilon_{ij}$	Randomized Complete Block (RCB)

The models all have G, the grand mean, and ε_{ij} in common and typically, we assume:

$\varepsilon_{ij} \sim N(0, \sigma^2)$.

This is the first place for an obvious change when considering non-normal data. The choice of the distribution, however, can influence the model itself.

How can distributions affect a statistical model?

Consider the Normal RCB case. The model, as shown above, will be:

$$g(Y_{ij}) = Y_{ij} = G + T_j + \beta_i + \varepsilon_{ij}$$

where ε_{ij} = the interaction of block and treatment effects, βT_{ij} .

When this model is applied to data such as the Binomial, however, it is considered incomplete. For Binomial, the mean should equal N*p. The mean would be estimated as $N*g^{-1}(G + T_j + \beta_i)$ and the variance would be computed as a function of the mean. The binomial mean, however, is defined by: $N*g^{-1}(G + T_j + \beta_i + \beta T_{ij})$. A piece of the mean, βT_{ij}, is missing if we assumed the Normal form of the model. Hence, both the mean and the variance estimates would be incomplete.

To obtain the correct estimation, we need to specify a fully saturated model as:

$$g(Y_{ij}) = G + T_j + \beta_i + \beta T_{ij} .$$

Note that, unlike the normal, there is no ε_{ij} term here. We do not require an extra term to estimate the variance of the binomial. Also, note that under the old PROC GLM LM framework this saturated model would produce errors and missing values. Under GLM and GLMM models, there are no “one case fits all” scenarios and care must be taken to formulate the statistical model for the assumed distribution.

More information on this topic can be found in:

1) *Stroup, W. W. 2014. Rethinking the Analysis of Non-Normal Data in Plant and Soil Science. Agron. J. 106:1–17. (Open Access)*
2) *Stroup, W. W. 2012. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Chapman & Hall/CRC Texts in Statistical Science, pp 555.*

Examples in SAS

The examples that follow are taken from Stroup 2014 (1) above. Example programs in SAS and R, as well as example data can be found in the supplementary materials of that article.

SAS code for binomial (count from a total data)

```
proc glimmix data=intro_binomial;
        class block Treatment;
        model Y=Treatment/dist=normal link=identity;
        random block;
        lsmeans Treatment / cl ilink;
run;
```

SAS code for binomial (count from a total data)

```
proc glimmix data=intro_binomial;
        class block Treatment;
        model Y/N=Treatment/dist=binomial link=logit;
        random block block*Treatment;
        lsmeans Treatment / cl ilink;
run;
```

The GLIMMIX procedure is similar to older procedures such as PROC GLM and PROC MIXED. There are still statements for CLASS, MODEL, RANDOM and LSMEANS. The options on the statements, however, differ to reflect the structure of GLMM model. The MODEL statement, for example, now has options to specify the distribution and its associated link with DIST= and LINK=, respectively. If these statements are omitted, SAS will assume a normal distribution with identity link. Note, however, that the binomial data is written in the model as a fraction, Y/N, representing the number of "positive" counts (Y) and the total number of counts (N). The binomial is the only distribution which uses this form and when GLIMMIX sees this in the model, it automatically assumes a binomial distribution.

The random statement works similarly to that of PROC MIXED, while the LSMEANS statement has two new options 'cl' and 'ilink'. The 'cl' specification produces confidence limits (95% by default) for the means. The 'ilink' option is very useful as it converts the means and confidence limits from the transformed Model Scale to the original Data Scale for reporting. Pair-wise mean comparisons can still be obtained through the DIFF or PDIFF options on LSMEANS.

Other types of data and model specifications can be specified for non-normal data. Count data, for example, can be modeled with the Poisson or the negative binomial distributions. The choice between these distributions depends on the nature of data and model diagnostics. As before, the distribution type will also determine the appropriate statistical model.

SAS code for Poisson (count type data)

```
proc glimmix data=intro_count;
        class block trt;
        model count=trt/d=poisson;
        random block block*treatment;
        lsmeans trt / cl ilink;
run;
```

SAS code for negative binomial (count type data)

```
proc glimmix data=intro_count;
        class block trt;
        model count=trt/d=negbin;
        random block;
        lsmeans trt / cl ilink;
run;
```

Here, the Poisson, like the binomial, uses the saturated model, while the negative binomial does not The distribution option can be abbreviated as'd='. SAS will also automatically pick the default link associated with the distribution if the 'LINK=' option is omitted. The choice of distributions depends on a condition referred to as *over dispersion*. This condition occurs when the data displays more variability than is expected from the specified model and distribution. Count data are often attributed to a Poisson process; however, many natural science problems show more variability than could be accounted for by the Poisson distribution. A recommended remedy is to use the negative binomial distribution, which has an additional scale parameter that can adjust for the "extra" variability. It should also be noted that over dispersion can occur with an incorrect model. Over dispersion can be diagnosed on the output using the reported statistic: 'Gener. Chi-Square / DF'. A value greater than one (>1.00) will indicate over dispersion. In this particular case, the Poisson produces a value of 1.08, which, while greater than 1.00, is still fairly moderate, not indicating much of an over dispersion. The corresponding value for the negative binomial would be 0.98, which is also good. The Poisson is probably adequate for this data and would be the more parsimonious choice (has the fewest parameters), although the negative binomial has a lower over dispersion statistic and produces the only significant treatment effect. Typically, the choice is not this close and when obvious over dispersion does occur, the negative binomial can usually account for the variability better and provide more statistical power for inference. Finally, note that the over dispersion diagnostic should not be used with the normal distribution.

In general, if the variance of the distribution involves parameters other than the mean, then proceed with the "typical model". Otherwise, the full model should be specified.

A third type of non-normal data is the continuous proportion. Such data often occur when measurements are taken relative to a standard (% of control) or as a percentage of a total amount (% ground cover, % area damaged, etc). While the data may be continuous, like normally distributed data, it will actually be different, in that it is restricted to the range 0.0 to 1.0 (0 to 100%). This type of data can often utilize a special distribution, the Beta distribution, which is naturally restricted between 0.0 and 1.0. It has the advantage of also displaying either the right or left skewness, two conditions that are hard to account for normal distribution.

SAS code for continuous proportion data using Normal (Don't use!!)

```
proc glimmix data=beta_intro method=quadrature;
        class block trt;
         model proportion=trt/d=normal;
        random intercept / subject=block;
         lsmeans trt / cl ilink;
        title 'glmm - proportion ~ Normal ';
run;
```

SAS code for continuous proportion data using Beta (Use this)

```
    proc glimmix data=beta_intro method=quadrature;
            class block trt;
            model proportion=trt/d=beta;
            random intercept / subject=block;
            lsmeans trt / cl ilink;
            title 'glmm - proportion ~ Beta ';
    run;
```

While both programs run fine, providing means, tests, and confidence intervals, a closer look at the results from the Normal distribution shows a problem. The confidence interval for the second treatment is 0.64 to 1.13, an impossible condition. The Beta distribution is a better option in this case, with results that are within the expectations, having an interval of 0.46 to 0.94.

Estimating statistical power and sample size

PROC GLIMMIX can also be used to compute power and sample size values. This is useful because the normal approximations typically used to estimate these quantities will not work well with non-normal data. As an example, we can look at the negative binomial count data and the binomial data examples given above. To compute power for these data, we rerun GLIMMIX, holding the specified parameters, fixed and random, at their estimated values. Next, we use the saved output of GMLMMIX in a "data step" to compute the power. For example, in the negative binomial count data above, the estimated means were Trt1: 5.9251 and Trt2: 22.3902, while the random effects for Blocks and scale were 0.6294 and 0.8664, respectively. The following code uses these values to compute estimated statistical power for 2 to 16 blocks and then prints and plots the results.

```
data NegBin;
    input Treatment count;
    do B = 2 to 16;
    do block = 1 to B;
        output;
    end;
    end;
datalines;
1   5.9251
2   22.3902
;
run;
proc sort;
    by B;
```

```
proc glimmix data=NegBin noprofile;
    class block Treatment;
    model count=Treatment/d=negbin;
    random block;
    parms (0.6294) (0.8664)/hold=1,2;
    ods output tests3=tests;
    by B;
run;

data power;
    set tests;
    alpha = 0.05;
    nc = numdf*fvalue;
    fcrit = finv(1 - alpha, numdf, dendf, 0);
    power = 1 - probf(fcrit, numdf, dendf, nc);
run;
proc print;
run;
symbol1 i = join v = dot c=black;

proc gplot;
    plot power*B=1;
run;
```

A similar process may be iteratively repeated at several settings for the binomial. For example, a common question with binomial data would be: *Is it better to have more blocks, or is it better to have more samples within a block?* The following code can be used for this type of exploration. It sets up an example data set with 2 to 16 blocks (B) and 10 to 300 samples per block (N) and then evaluates and plots the results for each combination.

```
data binomial_power;
    input Treatment P;
    do N = 10, 30, 50, 100, 200, 300;
        Y = N*P;
        do B = 2 to 16;
            do block = 1 to B;
                output;
            end;
        end;
    end;
cards;
0 .9276
1 .7807
;
```

```
proc sort;
	by N B;

proc glimmix data= binomial_power noprofile;
	class block Treatment;
	model Y/N= Treatment;
	random block block*Treatment;
	parms (0.5201) (0.8335)/hold=1,2;
	ods output tests3=tests;
	by N B;

data power;
	set tests;
	alpha = 0.05;
	nc = numdf*fvalue;
	fcrit = finv(1 - alpha, numdf, dendf, 0);
	power = 1 - probf(fcrit, numdf, dendf, nc);
proc print;
run;

symbol1 i = join v = dot c=black;
symbol2 i = join v = dot c=blue;
symbol3 i = join v = dot c=green;
symbol4 i = join v = dot c=red;
symbol5 i = join v = dot c=brown;
symbol6 i = join v = dot c=orange;

proc gplot;
	plot power*B=N;
run;
```

For more information, see: *Walter W. Stroup. 1999. Mixed Model Procedures To Assess Power, Precision, And Sample Size In The Design Of Experiments. (Available online at: http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms/Power/stroup-1999-power.pdf)*