STAT 5200 Handout #26

Generalized Linear Mixed Models

Up until now, we have assumed our error terms are normally distributed. What if normality is not realistic due to the nature of the data? (For example, what if the response variable isn't even continuous?) PROC GLIMMIX is flexible enough to allow for many common distributions, in all of the experimental designs we have discussed.

General notation: (recall matrix form on p. 3 of Handout #25)

$$\boldsymbol{\eta} = g(E[\mathbf{Y} \mid \boldsymbol{\gamma}]) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$$

($\mathbf{X}\boldsymbol{\beta}$: fixed effects; $\mathbf{Z}\boldsymbol{\gamma}$: random effects)

- g is called the 'link' function; there is usually a traditional link function for each common distribution of Y. This function links the actual response Y to the linear scale of effects ($\mu + A_i + B_j + \ldots$)

- For non-normal Y, just tell GLIMMIX the distribution and link function, and most everything else will run the same. Distributions can include normal, binomial, Poisson, gamma, exponential, negative binomial, Cauchy, F, …

Example: Researchers studied 16 varieties of wheat for their resistance to infestation by a certain pest. They arranged the varieties in a randomized complete block design, with four fields as blocks (**B**). Each Variety (**V**) is assigned to one section within each field. The outcome of interest was the number of damaged plants (**Dam**) out of the total number of plants (**Tot**) growing in the section.

```
data infestation;  input Block Variety Tot Dam @@; cards;
 1 14  8  7    1 16  9  7    1  7 13  9    1  6  9  9
 1 13 19  1    1 15 14  1    1  8  8  6    1  5 11  9
 1 11 12  2    1 12 11  8    1  2 10  8    1  3 12  5
 1 10  9  7    1  9 15  8    1  4 19  6    1  1  8  7
 2 15 15  1    2  3 11  9    2 10 12  5    2  2  9  9
 2 11 20 19    2  7 10  8    2 14 12 11    2  6 10  7
 2  5  8  8    2 13 16  1    2 12  9  2    2 16  9  0
 2  9 14  9    2  1 13 12    2  8 12  6    2  4 14  7
 3  7  7  7    3 13 17  1    3  8 13  3    3 14  9  0
 3  4 15 11    3 10  9  7    3  3 15 11    3  9 13  5
 3  6 16  9    3  1  8  8    3 15 17  1    3 12 12  8
 3 11  8  7    3 16 15  4    3  5 12 11    3  2 16 12
 4  9 15 10    4  4 10  8    4 12 13  7    4  1 15  9
 4 15 17  1    4  6  8  4    4 14 12 13    4  7 15 10
 4 13 18  1    4  8 13 11    4  3  9 11    4 10  6  8
 4  2 12 10    4 11  9 10    4  5 12 12    4 16 15 11
 ;
```

```
/**** Approach 1: Treat proportion as normal ****/

data infestation; set infestation;
 Prop = Dam/Tot;
proc glimmix data=infestation plots=residualpanel;
   class Block Variety;
   model Prop = Variety / dist=normal link=identity;
   random Block;
   title1 'Treating Proportion as Normal';
run;
```

defaults

Normal distribution and identify link function g(Y)=Y are default.

Treating Proportion as Normal

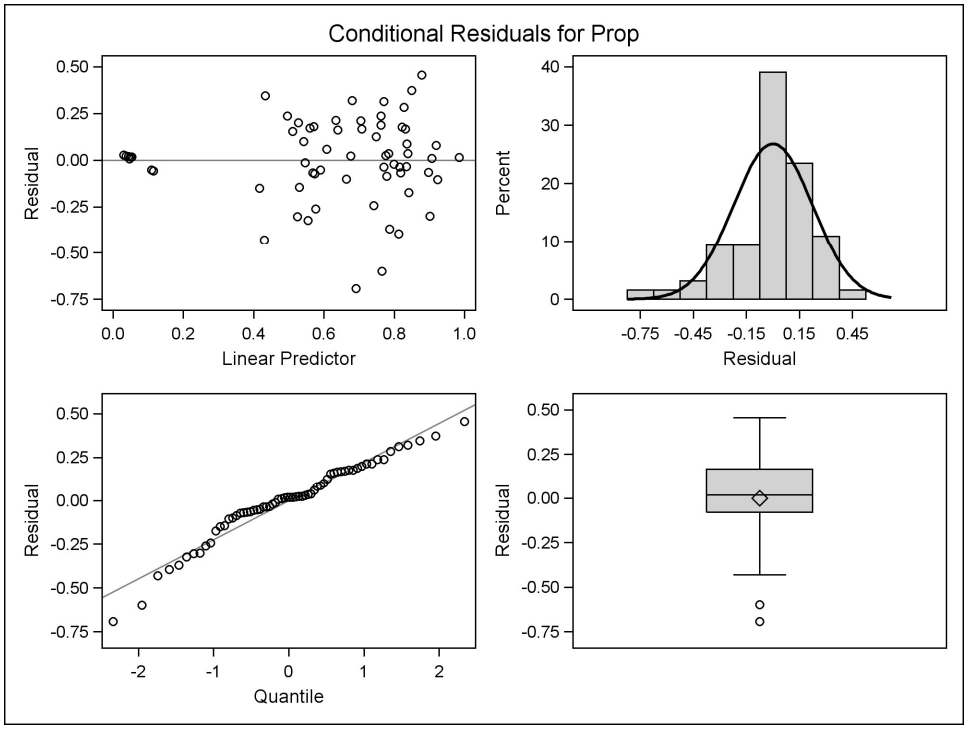
Convergence criterion (ABSGCONV=0.00001) satisfied.

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error
Block	0.002979	0.005924
Residual	0.06711	0.01415

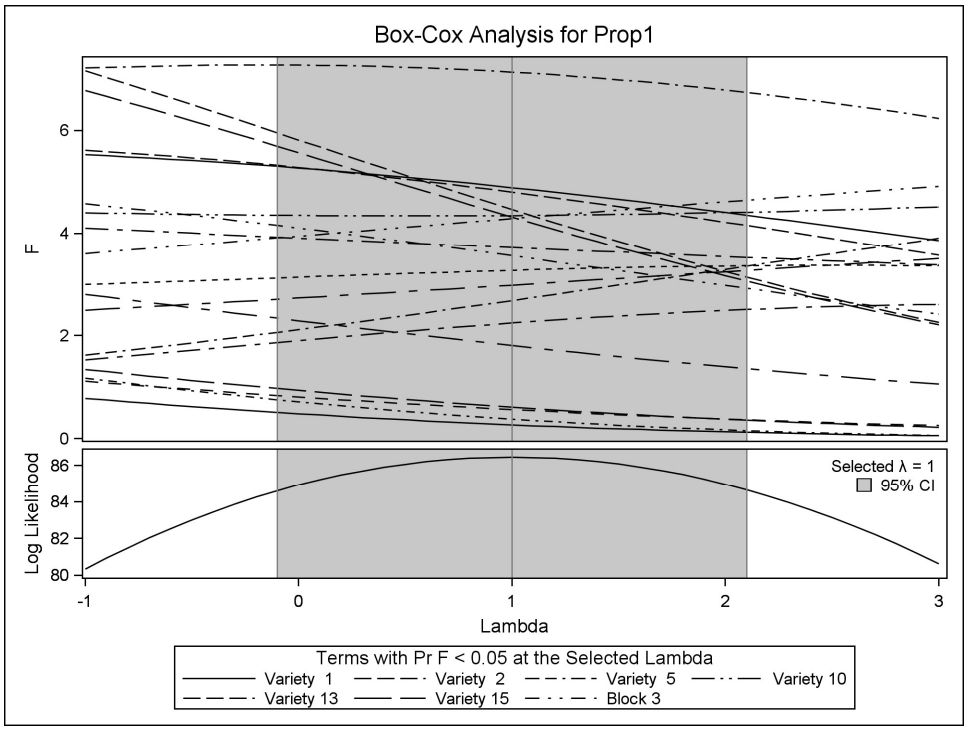
Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Variety	15	45	4.02	0.0001



```
/* Try Box-Cox transformation; use +1 to allow
   lambda=0 [log(Prop)] when Prop=0 */
data temp; set infestation;
  Prop1 = Prop+1;
proc transreg data=temp;
  model boxcox(Prop1 / lambda=-1 to 3 by 0.1)
        = class(Variety Block);
  title1 'Box-Cox on Proportion';
run;
```

Box-Cox on Proportion



While the proportion Dam/Tot may be considered continuous, normality may be problematic. Instead of assuming a Normal distribution for the data, note that a **Binomial distribution** may more appropriate:

$Y \sim \text{Binomial}(N, p)$

Y: # damaged; N: total # in sector; p: Prob. of damage

```
/**** Approach 2: Treat Dam as Binomial ****/
```

$$Dam \sim Binomial(Tot,p)$$

Here p is the probability of plant damage. That is, each plant's response is a binary Y (0 = no damage, 1 = damage), and p=P{Y=1}, or p=E[Y].

We are interested in modelling how p depends on the design factors. The "logit" transformation [g(p) below] is traditionally used to "link" this [range-restricted] probability scale to the [unrestricted] linear model scale:

$$g(p_{ij}) = log\frac{p_{ij}}{1-p_{ij}} \qquad \eta_{ij} = \mu + V_i + B_j$$

```
proc glimmix data=infestation;
   class Block Variety;
   model Dam/Tot = Variety / dist=binomial link=logit;
   random Block;
   title1 'Binomial Count Data';
run;
```

Binomial Count Data

Model Information	
Response Variable (Events)	Dam
Response Variable (Trials)	Tot
Response Distribution	Binomial
Link Function	Logit

Convergence criterion (PCONV=1.11022E-8) satisfied.

Fit Statistics	
-2 Res Log Pseudo-Likelihood	196.50
Generalized Chi-Square	123.31
Gener. Chi-Square / DF	2.80

The Chi-Square / DF ratio should be fairly close to 1. When it is clearly larger (as here), this suggests "overdispersion", meaning there is something causing variance that is not accounted for in the model.

Covariance Parameter Estimates		
Cov Parm	**Estimate**	**Standard Error**
Block	0.01007	0.03398

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Variety	15	41	7.89	<.0001

Overdispersion usually results from a mis-specified model, like omitted predictors, an incorrect link function, or an incorrect distributional assumption.

The Binomial approach above assumes each plant within a section is like an independent trial (like a toss of a coin, with a certain probability of heads [or damage here]). Maybe they are not independent (because each section [experimental unit] is like a subject), and dependence can cause overdispersion.

```
/**** Approach 3: Treat Dam as Poisson ****/
```

An alternative approach for count data is to use the **Poisson distribution**:

$$Dam \sim Poisson(\ Tot \times p\ )$$

Here p is still the probability of plant damage (on any given "attempt" [plant] within an experimental unit [section]). The "log" transformation [g(p) below] is traditionally used to "link" this [range-restricted] probability scale to the [unrestricted] linear model scale:

$$E[Dam] = Tot \cdot p$$

$$g(p_{ij}) = \log(p_{ij}) \qquad \eta_{ij} = \mu + V_i + B_j$$

$$\log\left(E\left[Dam_{ij}\right]\right) = \log\left(Tot_{ij}\right) + \log\left(p_{ij}\right) = \log\left(Tot_{ij}\right) + \mu + V_i + B_j$$

- This log link will lead to non-convergence in GLIMMIX if any levels of V or B have Dam=0 for all replicates, since log(0) is undefined. Log is natural log.

- The log(Tot) term in the linear model is called the "offset" and can be thought of as representing the "exposure" level in the experimental unit; can be defined inside PROC GLIMMIX. (This is not anything of interest, but it's critical to account for it.)

```
proc glimmix data=infestation;
  class Block Variety;
  logTot = log(Tot);
  model Dam = Variety / dist=poisson link=log
                        offset=logTot;
  random Block;
  lsmeans Variety / adjust=tukey lines plot;
  title1 'Poisson Count Data';
run;
```

Poisson Count Data

Model Information	
Response Variable	Dam
Response Distribution	Poisson
Link Function	Log
Offset Variable	logTot = log(Tot);

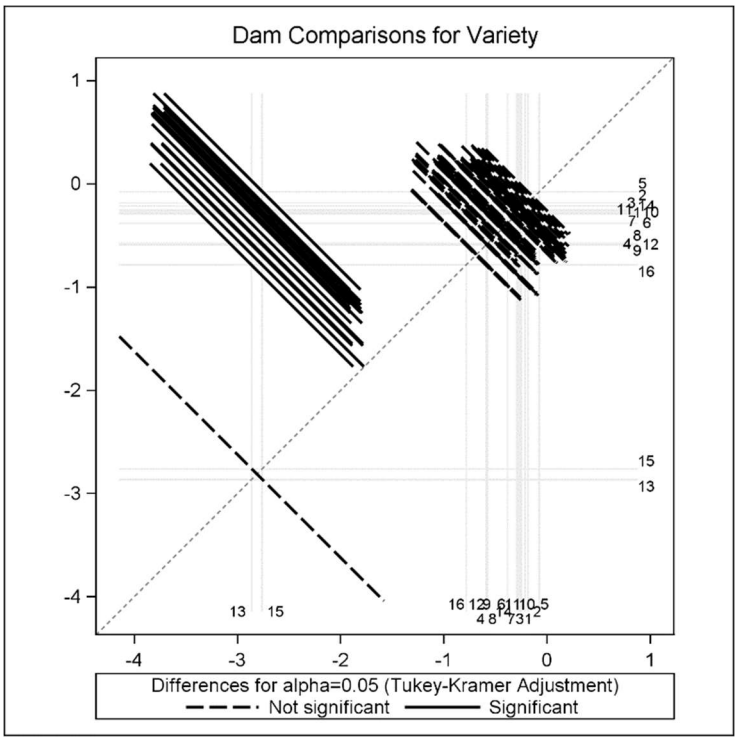
Convergence criterion (PCONV=1.11022E-8) satisfied.

Fit Statistics	
-2 Res Log Pseudo-Likelihood	78.38
Generalized Chi-Square	51.11
Gener. Chi-Square / DF	1.06

Covariance Parameter Estimates		
Cov Parm	**Estimate**	**Standard Error**
Block	0.007774	0.01333

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Variety	15	45	4.24	<.0001

Tukey-Kramer Grouping for Variety Least Squares Means (Alpha=0.05)		
LS-means with the same letter are not significantly different.		
Variety	**Estimate**	
5	-0.07500	A
…	…	A
16	-0.7865	A
15	-2.7600	B
13	-2.8639	B



Note: LSMEANS are on the response variable scale: η = log(p)
So the predicted probability of damage for Variety 15 is exp(-2.76) = 0.063.