Introduction to Generalized Linear Mixed Models

Analyzing Count Data

Jerry W Davis, Experimental Statistics, University of Georgia, Griffin Campus. 2018.

UNIVERSITY OF GEORGIA
Griffin Campus

Traditional Analysis of Variance

Analysis of Variance rests on three basic assumptions

- Response variables are normally distributed
- Individual observations are independent
- Variances between experimental units are homogeneous

Small deviations from these assumptions are not critical

- Analysis of Variance methods are very robust
- Central Limit Theorem says that data with many observations have normally distributed means

Analyzing non-normal data

Old techniques:

- Assumptions were / are often ignored
- Transformations were used to “normalize” the data

Improvements:

- Advances in statistical techniques allow categorical data to be modeled like normal data
- Computer hardware and software can solve numerically intensive problems
- Analysis of variance models incorporate different distributions so normality assumptions are unnecessary

Analysis of Variance Models

- Linear Models (LM), normal (Gaussian) data, PROCs GLM, REG, ANOVA
- Linear Mixed Models (LMM), normal (Gaussian) data, random and / or repeated effects, PROC MIXED
- Generalized Linear Models (GLM), non-normal data, PROCs LOGISTIC, GENMOD
- Generalized Linear Mixed Models (GLMM), normal or non-normal data, random and / or repeated effects, PROC GLIMMIX
- GLMM is the general model with LM, LMM and GLM being special cases of the general model

Generalized Models

- The term ***generalized*** refers to extending linear model theory to include categorical response data.

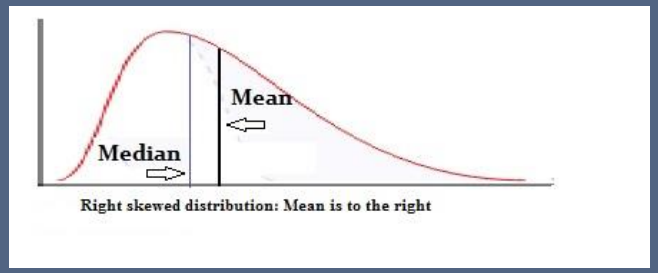
- Non-normal data can be analyzed in a conventional analysis of variance framework
  - F and *t* tests
  - Mean separation tests

Generalized Linear Mixed Models

- Selecting the correct distribution for categorical response variables is very important
- Follow the guidelines presented here and in the references
- Error messages provide little or no guidance in matching a response variable to a distribution
  - Never seen: Hey, your distribution does not match your response variable

Counts: Poisson or Negative Binomial distribution

- Non-negative integers, often right skewed



- Number of insects, weeds, or diseased plants, etc., within an experimental unit
- Counts are unbounded.
  - Biological limits (cotton bolls / plant) are not bounded – OK
  - The number of plants that died out of ten is bounded – not OK

Binomial: Binomial distribution

- Discrete positive integers between 0 and ***n***
- The number of successes from ***n*** independent trials
- When ***n*** equals 1, it is a Bernoulli trial (coin toss)
- Usual outcomes are 1 or 0, alive or dead, success or failure
- Also called discrete proportions - events per the number of trials
  - Eggs hatched from the total number of eggs
  - Seeds germinated from the total number of seeds planted
  - Plants that survived drought from the total number of plants

Continuous proportion: Beta distribution

- Proportion of affected area within an experimental unit
- Expressed as a percent (decimal fraction, i.e., 0.12, 0.25, etc.)
  - Area of a plot with disease or insect damage
  - Damaged leaf area
  - Lesion size per total area

Ratings and ranks: Multinomial distribution

- Subjective measurement based on a discrete scale or criteria
- Disease ratings, sensory evaluations, herbicide efficacy rating
- Response variables need not be numeric. Good, fair or poor are as valid as 1, 2 or 3.
- Linear scale implies the difference between ratings is equal
- Non-linear scale implies the differences are not equal
  - The difference between ratings 1 and 2 on a six point scale is not equal to the difference between ratings 5 and 6

Poisson distribution

- Mean is equal to the variance ($\mu = \sigma^2$)
- Implies that counts are uniformly distributed
- Agriculture counts tend to be clustered -> over-dispersed
- Over-dispersion – variance is larger than the mean
  - May cause inflated F values and underestimate standard errors
- Under-dispersion – variance is smaller than the mean
  - Less problematic

Negative Binomial distribution

- Similar to the Poisson distribution
- Includes a scale parameter (δ) so the mean and variance need not be equal ($\mu \neq \sigma^2$)
- May be more appropriate for agricultural counts

Pseudo – likelihoods

- Residual Pseudo-likelihood (RSPL)
- Default estimation method for GLIMMIX and non-normal data
- Does not produce a true log-likelihood

Consequences:

- Model is not conditioned by the random effects
- Only a conditional model can diagnose over-dispersion
- Fit statistics (AIC, BIC, AICC, etc.) cannot be calculated

Pseudo – likelihoods

Solution:

- Change estimation method to adaptive quadrature or Laplace
- Both methods fit a true log likelihood function
- Add **method=laplace** or **method=quad** to the PROC GLIMMIX statement
- Try one method and if there are problems switch to the other

Adaptive quadrature

- Adaptive quadrature is said to be more accurate than Laplace

Side effects:

- **ddfm** options **kr2** and **satterthwaite** are not available
  - Omit the **ddfm** option so GLIMMIX will default to **containment**
- Random effects must be processed by subjects
  - **random intercept / subject=block;**
  - **random block**; does not work

- Continued…

Adaptive quadrature

- When there are two random effects, such as block and year, writing two separate random statements is flagged as an error.

  For example:

  - ✗ random intercept / subject=block;
  - ✗ random intercept / subject=year;

  Solution: use one statement with the interaction term

  - ✓random intercept / subject=block*year;

Laplace

- Laplace is less restrictive than adaptive quadrature
- **ddfm=kr2** or **satterthwaite** is not available for Laplace
  - Omit the **ddfm** option so GLIMMIX will default to containment
- Random effects need not be processed by subjects, but it is a good idea to do so
- Multiple random statements are allowed
  - ✓ random intercept / subject=block;
  - ✓ random intercept / subject=year;
- The Laplace option may be a better first choice than quadrature
  - Less restrictive and often works better

Repeated Measures

- PROC GLIMMIX uses a random statement and the residual option to model repeated **(R-side)** effects.
- Adaptive quadrature and Laplace cannot model **R-side** effects
- Repeated effects must be modeled using random **(G-side)** effects
- Method is similar to doing a “split-plot in time”

- The difference is subtle and illustrated with an example

Link functions

- Response data remains on the original data scale when a model is fit
- LS-means are on the model scale when they are estimated
- The link function ***links*** the model scale means back to the data scale
- This is not the same as transforming the data, fitting a normal theory model and then back transforming the means
- Each distribution has a default link function

Distributions and Link functions

Distribution	Link Function	Syntax dist=	Syntax Link=
Beta	Logit	dist=beta	link=logit
Binomial	Logit	dist=binomial | bin | b	link=logit
Normal	Identity	dist=gaussian | g | normal | n	link=identity | id
Multinomial	Cumulative logit	dist=multinomial | multi | mult	link=cumlogit | clogit
Negative binomial	Log	dist=negbinomial | negbin | nb	link=log
Poisson	Log	dist=poisson | poi	link=log

Concepts for fitting a GLMM

- An analysis of variance model is a vector of linear predictors (equation) with unknown parameter estimates
- Every distribution has a corresponding likelihood function
- The vector of linear predictors is substituted into the likelihood function
- Parameters are estimated by minimizing the –log likelihood function
- LS-means are derived from the parameter estimates and are on the model scale
- The link function converts the model scale LS-mean estimates back to the original data scale

Key concepts

- PROC GLIMMIX uses a distribution to estimate model parameters
- PROC GLIMMIX does not ***fit*** the data to a distribution

- Response data values are not transformed by the link function
- The link function converts the LS-mean estimates back to the data scale after being estimated on the model scale

CountSeminar1.sas

Results – Negative Binomial

- Values for Pearson Chi-squared / DF and AICC are better than Poisson
- Accounted for over-dispersion
- Unbiased F-values and standard error estimates
- More intuitive LS-mean separation tests

- It may be better to start with a negative binomial distribution
  - I like to start with Poisson because it tells me if the data are over-dispersed, and may indicate or highlight other problems.

Results – Negative Binomial

- What if over-dispersion or other problems remain when using a negative binomial?
  - Examine the data for sparse or constant values.
  - Sparse data may need to be summed over treatment levels, locations or time
  - Some data may need to be sacrificed for a better analysis

- If all else fails, select the model with better diagnostics and interpretable results

Describing the analysis

The response data were analyzed with an analysis of variance model using a negative binomial distribution in PROC GLIMMIX (SAS/STAT, 2017).

Or something similar…

Repeated Measures analysis

- R-side model for normal data
  - random day / residual subject=rep*treatment type=cs;
  - ERROR: R-side random effects are not supported for METHOD=LAPLACE.

- G-side model for non-normal data method=quad or method=laplace
  - random day / subject=rep*treatment type=cs;
  - Remove the residual keyword and the repeated effect is modeled much like “split-plot in time”

  - See Appendix B for the full program

CountRepeatedSeminar1.sas

Repeated Measures analysis

- Other ways to handle repeated measurements
  - Analyze within the time or space variable
  - Sum across time or space
  - Summing has the potential advantage of handling sparse data

Common problems

- The zero covariance estimate for rep

Covariance Parameter Estimates			
Cov Parm	Subject	Estimate	Standard Error
Intercept	rep	0	.

- Causes this message

Estimated G matrix is not positive definite.

PoissonBadMeans.sas

Detecting problems

- Check the output tables to ensure that results are consistent
- Examine model diagnostic tables and the ANOVA table before the LS-mean tests
- If some of the results look fishy, something is probably wrong

- Never blindly accept computer output!

Detecting problems

- Sparse data is the most common problem I see when analyzing count data
- If nothing else can be done, delete treatment levels with zero or near zero counts
  - LS-means are always tested to see if they are different from zero
  - If a treatment level has all zero values, it is a constant
  - Therefore the initial test should be sufficient for a difference between the two
- Again, summing over time, space or treatment levels may be acceptable options when analyzing problem data
- Analyses within factors may also be an option

Detecting problems

- The problematic LS-mean tests can occur with GLM, MIXED and GLIMMIX
- Not limited to treatment levels with all zeros
- Other data problems may be the culprit
- Hard to diagnose problems from large multi-factor experiments with messy data
- Keep this in mind when designing the experiment

Convergence

- Mixed model procedures use iterative algorithms to minimize the –log likelihood function
- The algorithm (method) may encounter a problem and stop
- The process failed to *converge* and did not find a solution

- Non-convergence is like a lack-of-fit test – it implies that the data does not support the model

Convergence

- SAS messages indicating not-convergence
  - Note: An infinite likelihood is assumed in iteration 0 because of a non-positive definite R matrix for *'variable name' 'place'*
  - Note: Did not converge.
  - Error: Insufficient resources to determine number of quadrature points adaptively. The last successful…..

Convergence: causes and possible remedies

- Repeated measures – one observation per subject per time point
  - Possible data entry problem
- Response variable are large (10,000 – 1,000,000 range)
  - Divide values by a constant – will not affect significant tests
- Over parameterized model
  - Reduce the number of random effects
- Miss-specified model
  - Check the factors in the **class**, **model**, **random** and **repeated** statements

Convergence: causes and possible remedies

- Too many subjects
  - Combine subjects into groups, i.e., group individual cows into pens or lots
- Data problems
  - check data values carefully
- Change the maximum likelihood estimation method (**method=)**
- Maximum number of iterations reached
  - Increasing the number of iterations rarely works. Look for other problems.

Convergence: causes and possible remedies

- Sometimes simple remedies will not lead to convergence
  - Redefine the problem or scope of the experiment
  - Sacrifice some data to salvage tests for other factors
  - Do not get too creative with data manipulations
  - Be prepared to explain data or factor manipulations

- There is always PROC GLM…

Questions, Comments or Suggestions?

Jerry Davis

jwd@uga.edu

770 228-7237

UNIVERSITY OF GEORGIA
Griffin Campus