GLMM workshop 7 July 2016

Instructors: David Schneider, with Louis Charron, Devin Flawd, Kyle Millar, Anne St. Pierre Provencher, Sam Trueman

First session	1 PM	Room SN2109	Writing the model
Break			
Second session	2 PM	SN 2018/2025	F-ratios from Expected Mean Squares
Break			
Third session	3:30	SN 2067/2071	Executing the analysis

Goal of the first session – Writing Statistical Models

GLM	The General Linear Model	Fixed Effects + Normal Error
GzLM	The Generalized Linear model	Fixed Effects + Non-normal Errors
GLMM	The General Linear Mixed Model	Fixed + Random + Normal
GzLMM	The Generalized Linear Mixed Model	Fixed + Random + Non-normal

Goal of the second session - Writing out the expected mean squares
Forming unambiguous likelihood ratio tests (F, t, χ^2)

Goal of the third session - Executing a GLMM in a statistical package
Interpreting the output

First session 1 PM Room SN2109 Writing the model

Preliminaries

Definitions Nominal, Ordinal, Interval, and Ratio scale variables.

Definitions:

	GLM	General Linear Model
	GzLM	Generalized Linear Model
	GLMM	General Linear Mixed Model
	GzLMM	Generalized Linear Mixed Model

Series of examples to work through.

- Distinguish response from explanatory variables
- Assign symbols to all variables
  - Notational conventions Nominal scale variable ALL UPPER CASE
    - Ratio scale variables Begin with upper case.
    - β for fixed effect coefficients (slopes and contrasts)
    - μ for random effect parameters
- Write the model, calculate the df, complete the first 2 columns of the ANOVA table

Preview

GLM - Fixed Effects

- Single explanatory variable – 3 examples
  - Write the GLM Fixed Effect model
  - Write the degrees of freedom below each term in the model - - > Source df table
- Two explanatory variables – Crossed - 3 examples
  - Write the GLM – Fixed * Fixed - - > Source df table — Factor * Factor; Factor * Covariate; Covariate * Covariate

GzLM - Fixed Effects. The first solution to heterogeneous errors - 2 examples

GLM - Random Effects. The second solution to heterogeneous errors.

- Definition of Random Effects, Random variables.
- Identify explanatory variables as Random or Fixed
  - β Notation for fixed factors. μ notation for random factors
- Single explanatory variable - 1 example
  - Write the GLMM –- - > Source df table — Fixed * Random Effects
- Two explanatory variables – Nested - 1 example
  - How to distinguish nested from crossed factors
    - Write the GLM - - > Source df table — Random(Random)
    - Write the GLMM - - > Source df table — Fixed + Random(Random)

GLMM - Mixed Effects (Fixed and Random)

Two explanatory variables	Nested Example
	Crossed Example
Three explanatory variables	Nested Example
	Crossed Example (Latin Square)

GzLMM – Non-normal error and mixed effects (fixed and random factors)

GLM with a single fixed explanatory variable 3 examples.

Write the Fixed Effect GLM, calculate df, fill in the blank columns of the ANOVA table.

1. Pea section growth data, from Box 9.4 in Sokal and Rohlf (1995).
Does length depend on treatment (control versus 4 different sugars with auxin present) ?
10 measurements of length of pea section in each treatment group

Length	*Len*	Response variable, ratio scale
Treatment	*TRT*	Categorical explanatory variable

Sketch graph of response vs explanatory

Write the model $Len = \beta_o + \beta_{Trt}\, TRT + \varepsilon_{Normal}$

Calculate df (10*5) = 1 + (5-1) + 45

df total = ntot -1 TRT df = number of categories – 1

Source	df
TRT	
error	
total	49

Fill out first 2 columns of ANOVA table from model
http://www.mun.ca/biology/schneider/b4605/LNotes/Pt3/Ch10_3.pdf
http://www.mun.ca/biology/schneider/b4605/GLMMworkshop/Data/PeaSections.csv

2. Example 9.3.1 from Snedecor and Cochran (1989). Quantity of interest is the phosphorus content of corn (*Pcorn* in ppm), in relation to the phosphorus levels in samples of soils with experimentally fixed levels of phosphorus (*Psoil* in ppm). Does the phosphorus content of corn increase when organic soil phosphorus is increased ? *Pcorn* and *Psoil* are both ratio scale variables. 9 measurements of *Pcorn*, matched with 9 of *Psoil*

Model ______________________________

df ______________________________

Sketch graph of response vs explanatory

http://www.mun.ca/biology/schneider/b4605/LNotes/Pt3/Ch9_1.pdf

http://www.mun.ca/biology/schneider/b4605/GLMMworkshop/Data/PCorn.csv

Source	df
error	7
total	

GLM with a single fixed explanatory variable 3rd example.

3. Does inversion heterozygosity (HZYG) change with elevation above sea level (Hsl) in *Drosophila pseudoobscura*). Data are from Dobzhansky (1948) as reported in Brussard (1984).
One measurement of HZYG at each of 7 different elevations.

Source	df
	5

Response variable with symbol ______________________________

Explanatory variable with symbol ______________________________

Model ____

df ____

GLM with a single fixed explanatory variable Review

Definition of fixed effects:

1. *TRT* is a fixed effect because we are interested in the contrast among the 5 means.
   β_{TRT} is a set of unknown fixed effect contrasts.
2. *Psoil* is a fixed effect because we are interested in rate of increase in *Pcorn* with increase in *Psoil*.
   β_{Psoil} is the unknown rate.
3. *Hsl* is a fixed effect because we are interested in the whether *Hzyg* changes with elevation (altitude above sea level)
   β_{Hsl} is the fixed effect rate. $\hat{\beta}_{Hsl}$ is an estimate of β_{Hsl}

GLM with two fixed explanatory variables 3 examples Factor * Factor
Factor * Covariate
Covariate * Covariate

Format for writing a model with two explanatory variables

$Response = \beta_o + \beta_{V1} V1 + \beta_{V2} V2 + \beta_{V1\times V2} V1 \times V2 + \varepsilon_{Normal}$

The interactive term is written as the product of two component variables $\beta_{V1\times V2} V1 \times V2$

Verbal statement: The effect of V1 on the response variable depends on V2

Write the Fixed factor × Fixed factor GLM, calculate df, fill out the Source df table

df total = *ntot*-1 df $V1 \times V2$ = df($V1$) × df($V2$)

4. Does oxygen consumption VO_2 depend on salinity (100% 75% and 50% seawater) in two species of limpet (*Acmea digitalis* and *A. scabra)*? Eight measurements at 3 different salinities in each of two species *ntot* = 48. Data from Sokal and Rohlf (1995).

Response variable with symbol ______________________________

Explanatory variable	Symbol	Categorical or Ratio scale
__________________	______	___________
__________________	______	___________

Model ____

df ____

Source	df
	42

Interpret the interactive effect (state this in words)

http://www.mun.ca/biology/schneider/b4605/LNotes/Pt4/Ch13_1.pdf

http://www.mun.ca/biology/schneider/b4605/GLMMworkshop/Data/Limpets.csv

GLM with two fixed explanatory variables

Factor * Factor
2nd example -> Factor * Covariate (aka ANCOVA)
Covariate * Covariate

5. Does inversion heterozygosity (*Hzyg*) change with elevation above sea level (*Hsl*), in 2 species of *Drosophila* (SP = *D. persimilis* or *D. pseudoobscura*). Data are from Dobzhansky (1948) as reported in Brussard (1984). One measurement in each species at 7 different elevations.

Source	df
	10

Model ____

df ____________________________________

Complete the Source df table.
Interpret the interactive effect (state it in words)

http://www.mun.ca/biology/schneider/b4605/LNotes/Pt4/Ch14_1.pdf

http://www.mun.ca/biology/schneider/b4605/GLMMworkshop/Data/Brussard.csv

GLM with two fixed explanatory variables

Factor * Factor
Factor * Covariate
3rd example -> Covariate * Covariate (aka multiple regression)

6. Data from Snedecor and Cochrane 1980 Table 17.2.1
Does plant available phosphorus content of corn (ppm) from 17 Iowa soils at 20 deg C depend on inorganic and organic phosphorus in the soil?
.

Source	df
	13

Model ____

df ____

Complete the Source df table.
Interpret the interactive effect (state it in words)

http://www.mun.ca/biology/schneider/b4605/LNotes/Pt4/Ch12_1.pdf

http://www.mun.ca/biology/schneider/b4605/GLMMworkshop/Data/PAvailable.csv

GzLM- The first solution to heterogeneous errors - 2 examples (Poisson and Binomial)

The GLM applies to a ratio scale response variable *Y* with normal error ε_{Normal}

Count data usually violate the assumption of homogeneous residuals.

Ratio scale counts (counts in defined units, ranging from zero upward)

Use Poisson or Negative Binomial error model $\varepsilon_{Poisson}$ or $\varepsilon_{NegBinomial}$

So we use the Generalized Linear Model, which allows us to use a better error model.

The GzLM (which includes the GLM as a special case) has three components

1. The structural model consisting of linear predictors.

   For the GLM, the linear predictor is the sum of fixed factors and covariates.

   The linear predictor ANCOVA example (Brussard) was $\eta = \beta_o + \beta_{SP}SP + \beta_{Hsl}Hsl + \beta_{SP \cdot Hsl}SP \cdot Hsl$

2. A linkfunction, that links the linear predictor to the response variable.
3. The error

Model equation form: $Y = \beta_o + \beta_x X + \varepsilon_{Normal}$

Probability distribution form: $Y \sim Normal(\beta_o + \beta_x X, \sigma^2)$

This is read as : *Y* is normally distributed, given the parameters β_o , β_x , and σ^2 (the fixed variance)

The distributional assumption, given the parameters, can only be checked after estimating the parameters

Count data usually violate this assumption.-- > heterogeneous residuals

So we use a better error model (Generalized Linear Model)

Ratio scale counts (counts in defined units, ranging from zero upward)

Use Poisson or Negative Binomial error model $\varepsilon_{Poisson}$ or $\varepsilon_{NegBinomial}$

$$Count = e^{\eta} + \varepsilon_{Poisson}$$
$$\eta = \beta_o + \beta_{V1}V1 + \dots$$

Corps	Deaths
Guard	16
First	16
Second	12
Third	12

8. Death by horsekick. The classic example of Poisson data is the number of deaths by horse kick for each of 16 corps in the Prussian army, from 1875 to 1894. Bortkiewicz (1898 *The Law of Small Numbers*) showed that the horsekick data fit a Poisson distribution.

Symbol for response variable ___________ and for explanatory variable __________

Write the model $Odds = e^{\eta} + \varepsilon_{Poisson}$ (Fit to 1:1:1:1 assumes Poisson error)

$\eta =$ ___________________________

Likelihood Ratio Test: ΔG = 1.147 df = 3 p = 0.7658 Therefore, cannot reject 1:1:1:1 fit

http://www.mun.ca/biology/schneider/b4605/LNotes/Pt5/Ch17_2.pdf

GzLM **with fixed explanatory variables.** - 2nd example

The GLM assumes a normal error with fixed (constant variance) = ε_{Normal}

Count data usually violate this assumption.-- > heterogeneous residuals

So we use a better error model (Generalized Linear Model)

Nominal scale counts (units scored Y or N) Use binomial error model $\varepsilon_{binomial}$

Yes/No = *Odds* $Odds = e^{\eta} + \varepsilon_{binomial}$

$$\eta = \beta_o + \beta_{V1}V1 + \ldots$$

The response variable, Odds, are calculated as p/(1-p), where p is the ratio of success to number of trials.

9. Example – Cancer in cigarette smokers. Data from Cornfield (1951) who established the mathematical basis for using case-control samples to estimate risk in a population.

	Lung Tumors		
	Present	Absent	Total
heavy smokers	27	99	126
light smokers	8	72	80

Odds of tumor for

Heavy smokers ___________

Light smokers _____________

Odds ratio, heavy relative to light ___________

Symbol for response variable ____________ and for explanatory variable ___________

Write the model $Odds = e^{\eta} + \varepsilon_{binomial}$

(contingency test not correct. it assumes Poisson error instead of binomial)

η = _____________________________

The 95% confidence limits are 1.05 to 5.1.

The null hypothesis is OR = 1. Odds the same for light and heavy smokers)

Do the confidence limits exclude the null? ___________________________

http://www.mun.ca/biology/schneider/b4605/LNotes/Pt5/Ch18_3.pdf

GLM - Random Effects. The second solution to heterogeneous errors.

The GLM assumes a normal error with fixed (constant) variance $= \varepsilon_{Normal}$
Grouped data usually violate this assumption.-- > heterogeneous residuals
Examples: Paired data, clustered data, blocked data
Examples: Repeated measures (*e.g.* 3 samples at one time), longitudinal data (3 samples in sequence)
To capture this heterogeneity, we introduce a random effect variable *Z* with random coefficients τ (tau).

$$Y = \mu_o + \tau_Z Z + \varepsilon_{Normal}$$

μ_o = random intercept
$\tau_Z Z$ = random effectd

GLM Single Random Factor

10 The first published ANOVA table was Example 38 in Fisher (1925) *Statistical Methods for Research Workers*. "In an experiment on the accuracy of counting soil bacteria, a soil sample was divided into four parallel samples and from each of these after dilution seven plates were inoculated. The number of colonies on each plate is shown below in example 12 (Table 41). Do the results from the four samples agree within the limits of random sampling? In other words, is the whole set of 28 values homogeneous, or is there any perceptible intraclass correlation?"

Table 42	Degrees of Freedom	Sum of Squares	Mean Square	F-ratio	R^2	Likelihood Ratio
Between Classes (Soil sample)	3	1446				
Within Classes (Error)	24	94.96				

Assign a symbol to the response variable _____ and explanatory variable _________

Write the model (use μ and τ) ____

Compute both mean squares (= SS/df) and place them in the ANOVA table
Compute the ratio of the two means squares (the F-ratio) and place it in the table
Compute the explained variance R^2 = Between class SS/SS_{total} = _________________
Do the 4 samples deviate from random sampling? To find out we calculate the likelihood ratio.

$LR = (1-R^2)^{-n/2}$ = ________________

Likelihood Ratio test: Compare the F-ratio to the 5% p-value of the F-distribution
The 5% probability for the F-distribution (excel code) is: FINV(0.05,3,24) = 3.009
Do the results from the four samples agree within the limits of random sampling? _______

http://www.mun.ca/biology/schneider/b4605/GLMMworkshop/Data/FisherEx38.csv

GLM with two random factors 2 examples

Nested - Random within Random
Crossed - Random × Random

$$Y = \mu_o + \Sigma\tau_Z Z + \varepsilon_{Normal}$$
$$\Sigma\tau_Z Z = \text{sum of random effects of random variable } Z$$

11. Winglength of 12 mosquitos (3 cages, 4 flies per cage). The left wing of each fly was measured twice.

Source	df	SS	MS	F	---->	p
Cage	2	665.68	332.84	1.74		0.23
Fly⊂Cage	9	1720.68	191.19	147.07		<0.0001
Error	12	15.62	1.3017			
Total	23	2401.97				

ANOVA table
Table 10.1 in Sokal and Rohlf (1995).

Write the model from the Source and df columns in the ANOVA table

____

Show how each df was calculated: 2 = ________ 9 = ___________

23 = __________ 12 = __________

Note that the Cage F-ratio was not calculated with respect to the MS error.
The Cage F-ratio was calculated from a random factor, Fly(Cage). Why ? Stay tuned.

http://www.mun.ca/biology/schneider/b4605/LNotes/Pt4/Ch13_6.pdf

http://www.mun.ca/biology/schneider/b4605/GLMMworkshop/Data/FisherEx38.csv

GLM with two random factors

Nested - Random within Random

2nd example - - > Crossed - Random × Random

12. Fisher's Table 42 (Example 38) shows a nested design.

Plate	Sample			
	I	II	III	IV
1	72	74	78	69
2	69	72	74	67
3	63	70	70	66
4	59	69	58	64
5	59	66	58	62
6	53	58	56	58
7	51	52	56	54
Total	426	461	450	440
Mean	60.86	65.86	64.29	62.86

It ignores the fact that each plate was inoculated with subsamples from each of the four initial samples (Classes). Consequently, we can treat class (*i.e.* sample) as a random factor with 4 levels and cross it with another random factor, plate.
Assign symbols to both explanatory variables and write a two way random effects GLM with an interaction term.

Symbols ____

Model ____

Complete the Source and df columns of the ANOVA table for this model.
The correct model is a saturated model, the error term will have zero degrees of freedom.
We'll use this in the next session.

GLMM with two explanatory variables 2 examples Fixed + Random
Fixed × Random

The GLM assumes a normal error with fixed (constant) variance $= \varepsilon_{Normal}$
Grouped data often violate this assumption.-- > heterogeneous residuals
Paired data, clustered data, blocked data
Repeated measures (e.g. 3 samples at once), longitudinal data (3 sequential samples)
To capture this heterogeneity, we write a General Linear Mixed Model, which has both fixed and random effects.

$$Y = \beta_o + \Sigma\beta_X X + \Sigma\tau_Z Z + \varepsilon_{Normal}$$
$$\Sigma\beta_X X = \text{sum of fixed effects}$$
$$\Sigma\tau_Z Z = \text{sum of heterogeneous random effects}$$
$$\varepsilon_{Normal} = \text{homogeneous normal errors}$$

<u>Random or Fixed?</u> The definition of fixed versus random differs among text books.
Definition from Quinn and Keough (2002)
There are two types of categorical predictor variables in linear models. The most common type is a <u>fixed factor</u>, where all the levels of the factor (*i.e.* all the groups or treatments) that are of interest are included in the analysis. We cannot extrapolate our statistical conclusions beyond these specific levels to other groups or treatments not in the study. If we repeated the study, we would usually use the same levels of the fixed factor again. Linear models based on fixed categorical predictor variables (fixed factors) are termed fixed effects models (or Model 1 ANOVAs). Fixed effect models are analogous to linear regression models where X is assumed to be fixed.
The other type of factor is a <u>random factor</u>, where we are only using a random selection of all the possible levels (or groups) of the factor and we usually wish to make inferences about all the possible groups from our sample of groups. If we repeated the study, we would usually take another sample of groups from the population of possible groups.

Drawing a branching tree diagram is not a reliable way to distinguish crossed from nested designs.
Why? Because a crossed design can be drawn as a branching tree.
The reliable way to distinguish crossed and nested designs is to write all of the two way tables and fill in the sample size in each cell of each table. If all (or most) of the cells have at least one sample then the two variables are crossed. If not the two factors are nested. For three factors there are three pairs and so three two-way tables.

13. Wheat Yields from Cornell (1971)

Three pots were assigned to each treatment.

The two-way (Pot × Treatment) table now has 12 cells.

There is 1 sample in each cell.

When we do the cross test the design appears to be crossed.

However, there were 12 pots in the experiment, not 3.

Treatment	Pot Number	Plant number 1	2	3
None	1	20.6	22.3	19.8
None	2	23.4	21.9	22.8
None	3	21.8	20.6	21.3
Straw	1	13.6	13.9	14.2
Straw	2	13.7	14.5	13.8
Straw	3	12.9	13.1	13.4
Straw + PO4	1	14.8	14.6	14.9
Straw + PO4	2	14.3	13.9	13.5
Straw + PO4	3	14.4	13.8	14.1
Straw+PO4+lime	1	14.1	13.8	14.3
Straw+PO4+lime	2	14.0	13.9	14.2
Straw+PO4+lime	3	14.4	14.1	13.6

http://www.mun.ca/biology/schneider/b4605/GLMMworkshop/Data/WheatYield.csv

Recode the Pot variable to show that there are 12 pots.

The two-way (Pot × Treatment) table now has 36 cells.

Most of the cells are empty.

We cannot estimate Pot × Treatment.

Pot is nested within treatment Pot(Treatment)

Carry out the cross test for Pot × Plant and Trt × Plant.

Now many cells? ____________________

How many empty cells? __________________

Can Pot × Plant be estimated ? Y/N _______________

Can Trt × Plant be estimated ? Y/N _______________

Treatment	Pot Number	Plant number 1	2	3
None	1	20.6	22.3	19.8
None	2	23.4	21.9	22.8
None	3	21.8	20.6	21.3
Straw	4	13.6	13.9	14.2
Straw	5	13.7	14.5	13.8
Straw	6	12.9	13.1	13.4
Straw + PO4	7	14.8	14.6	14.9
Straw + PO4	8	14.3	13.9	13.5
Straw + PO4	9	14.4	13.8	14.1
Straw+PO4+lime	10	14.1	13.8	14.3
Straw+PO4+lime	11	14.0	13.9	14.2
Straw+PO4+lime	12	14.4	14.1	13.6

GLMM with two explanatory variables 2nd example Fixed × Random

Subject	Drug A	Drug B
1	0.7	1.9
2	-1.6	0.8
3	-0.2	1.1
4	-1.2	0.1
5	-0.1	-0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4

14. Sleep data (Cushny and Peebles), used by Student (W. Gossett) to introduce the *t*-test. Data are: hours of extra sleep with two drugs Hyoscyamine (Drug A) and L Hyoscine (Drug B), each administered to 10 subjects. Values reported are averages. The pairing across subject allows us to remove the effects of individual variation.

Assign a symbol to the response variable ________________

For each explanatory variable assign a symbol and state reason for assigning it as Fixed or Random

________ ____

________ ____

http://www.mun.ca/biology/schneider/b4605/LNotes/Pt4/Ch13_3.pdf

http://www.mun.ca/biology/schneider/b4605/GLMMworkshop/Data/ExtraSleep.csv

Crossed or Nested?

There are only two variables, hence only one interaction term.
We can see right away that this is a crossed design.