

Introduction to Generalized Linear Mixed Models (GLMMs)

Front Range Student Ecology Symposium 2025

Dr. Ann Hess, CSU Stat Lab

Introduction

Linear Models (LM)

Examples: Monarch and Bat data

Generalized Linear Models (GLM)

Mixed Models (LMM, GLMM)

Conclusions

References and Resources

Introduction

- Appropriate statistical analysis and presentation of results an important part of research.
- This workshop will use two ecology data sets to motivate and introduce GLMMs.
- We will start with a brief review of linear models (LM).
- Next we will introduce generalized linear models (GLM) and mixed models (LMM).
- Finally we will put everything together with generalized linear mixed models (GLMM).

Linear Models (LM)

- Linear models (or regression models) are one of the most commonly used statistical methods!
- In R, we fit these models using `lm()`.
- Linear models require a numeric response, but can include numeric and categorical predictors.
- This class of models include simple linear regression, multiple regression, one-way ANOVA, multi-factor ANOVA models and beyond.

Linear Models (LM): Assumptions

Linear models have several assumptions.

Assumption	LM
Independence	Yes
Distribution	Normal
Variance	Equal

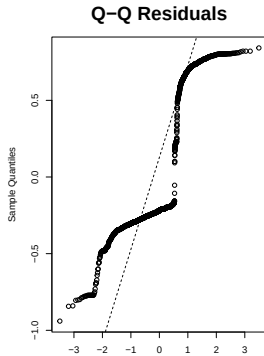
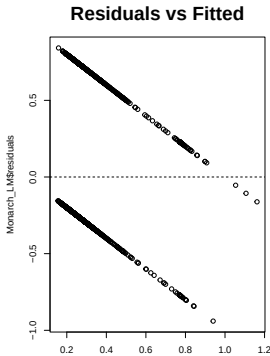
- To evaluate the assumption of independence, we consider the study design.
- To evaluate the assumptions of normality and equal variance, we can examine the residual diagnostic plots.
- Residuals capture the difference between the observed response (y) vs the predicted response ($\hat{y}$).

Monarch Data

- Thanks to Laura Lukens and Maggie Church for providing this data!
- This Citizen Scientist project monitored monarch butterfly activity across the US.
- This data includes a total of 2109 observations from 285 sites.
- Response:
 - **monarchpresence**: presence (1) or absence (0) of monarchs (binary)
 - Overall, monarchs were observed for 30% of observations.
- Predictors:
 - **mwcount**: number of milkweed stems
 - **year**: 2015-2024
 - **region**: N or S
- For this example, we consider only plots that included milkweed species *Asclepias tuberosa*.

Monarch Data: Why can't we use linear models?

- Linear models assume that the residuals are normally distributed.
- This assumption is NOT satisfied when we have a binary (0/1) response. See residual diagnostic plots below.
- In addition, a linear model could give predicted probabilities outside of (0,1).

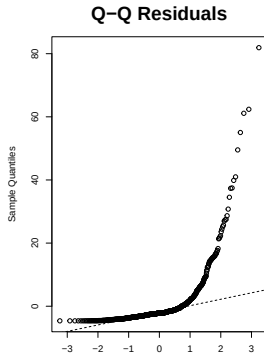
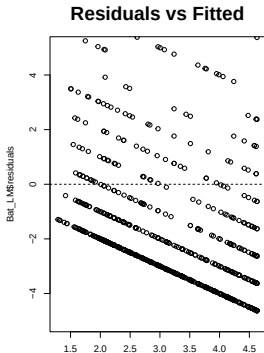


Bat Data

- From Cole (2022), bat activity was monitored in Grant Teton National Park via echo location.
- This data includes a total of 840 observations from 36 sites.
- Response:
 - **batcount**: number of bat calls
 - Overall, the average count was 3.18. The majority of counts were between 0 - 3.
- Predictors:
 - **brightscore**: summed brightness score of lights within 50 m
 - **moonpct**: lunar phase (0-100%)
 - **year**: 2016 or 2017
- For this example, we focus on species Epfu.

Bat Data: Why can't we use linear models?

- Linear models assume that the residuals are normally distributed.
- This assumption is NOT satisfied when we have a (low) count responses. See residual diagnostic plots below.
- In addition, a linear model could give predicted counts less than 0.



Generalized Linear Models (GLM): binomial family

- To specify a generalized linear model (GLM), we specify a family.
- For **binary** responses, we use **family = binomial**.
- This is typically called **logistic regression**.
- Monarch presence/absence is a binary response.
- In R, `glm(monarchpresence ~ mwcount, family = binomial)`

Generalized Linear Models (GLM): poisson or nbinom family

- For **count** responses, we use **family = poisson** or **family = nbinom**.
- Bat counts are a count response.
- In R, `glm(batcount ~ moonpct, family = poisson)`

Generalized Linear Models (GLM): other families

- The GLM models listed here are by far the most common, but other options are briefly mentioned later.
- If we do NOT specify a family, then `family = gaussian` is used by default. This brings us back to a linear model, with equivalent results to `lm()`.

Generalized Linear Models (GLM): Assumptions

We compare the assumptions for LM vs GLM.

Assumption	LM	GLM
Independence	Yes	Yes
Family	normal/ gaussian	binomial poisson or nb others
Variance	Equal	from family

- For binomial and poisson families, the variance is “dictated” by the distribution.
- Negative binomial is a more flexible extension of poisson that allows “over-dispersion”. Negative binomial is common in ecology!

Generalized Linear Models (GLM): Link functions

For each GLM family, there is a default link function.

Family	Link
binomial	logit
poisson	log
nbinom	log

- Other link functions are possible, but will not be discussed in this talk.
- We need to be aware of the link function when interpreting the results of a GLM model.

Monarch: Logistic Regression

- Logistic regression refers to a GLM with family = binomial (link = “logit”).
- We are modeling the probability ($\hat{p}$) of observing a monarch as a function of mwcount. For example, $\hat{p} = 0.10$ or 0.55 .

```
Monarch_GLM <- glm(monarchpresence ~ mwcount,  
                    family = binomial, data = MonarchData)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-1.211	0.058	-20.758	0
mwcount	0.012	0.001	11.016	0

Monarch: Logistic Regression Scales of Interpretation

- Logit = log odds scale (default for output, but not for presentation):

$$\log\left(\frac{\hat{p}}{1 - \hat{p}}\right) = -1.211 + 0.012 * \text{mwcout}$$

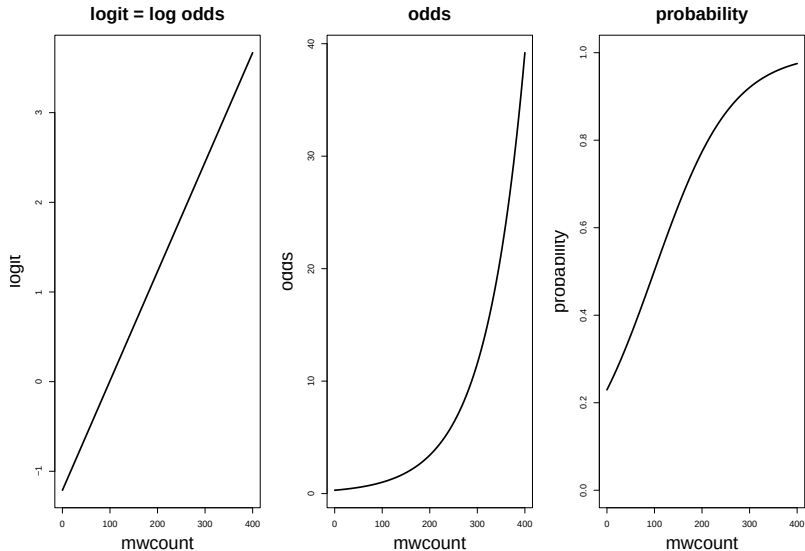
- Odds scale (most common for presentation):

$$\frac{\hat{p}}{1 - \hat{p}} = \exp(-1.211 + 0.012 * \text{mwcout})$$

- Probability scale:

$$\hat{p} = \frac{\exp(-1.211 + 0.012 * \text{mwcout})}{1 + \exp(-1.211 + 0.012 * \text{mwcout})}$$

Monarch: Logistic Regression Visualization



Monarch: Logistic Regression Interpretation

term	estimate	std.error	statistic	p.value
(Intercept)	-1.211	0.058	-20.758	0
mwcount	0.012	0.001	11.016	0

- Logit scale interpretation: An increase of 1 milk weed stem is associated with an (**additive**) increase of 0.012 in the estimated **log odds** of observing a monarch.
- Odds scale interpretation: An increase in 1 milk weed stem is associated with a **multiplicative** increase of $\exp(0.012) = 1.012$ in the estimated **odds** of observing a monarch.
- The value $1.012 (= \exp(0.012))$ is called the odds ratio.

Bat: Poisson Regression

We are modeling bat count as a function of moonpct.

```
Bat_GLM <- glm(batcount ~ moonpct,  
               family = poisson, data = BatData)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1.531	0.029	52.852	0
moonpct	-0.009	0.001	-15.225	0

For convenience, I use a poisson model here. The discussion would be very similar if a negative binomial model had been used.

Bat: Poisson Regression Scales of Interpretation

- Default Scale:

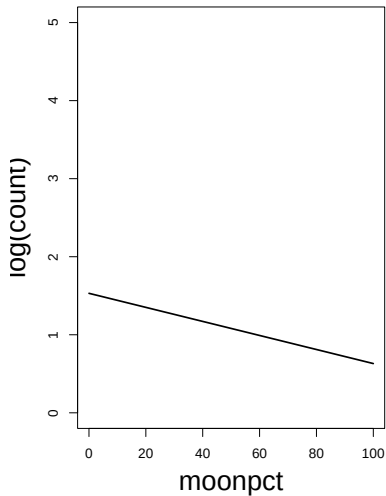
$$\log(\text{batcount}) = 1.531 - 0.009 * \text{moonpct}$$

- “Back Transformed” Scale:

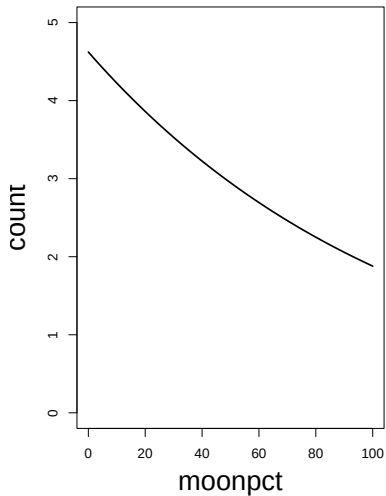
$$\text{batcount} = \exp(1.531 - 0.009 * \text{moonpct})$$

Bat: Poisson Regression Visualization

Original Scale



Back Transformed



Bat: Poisson Regression Interpretation

term	estimate	std.error	statistic	p.value
(Intercept)	1.531	0.029	52.852	0
moonpct	-0.009	0.001	-15.225	0

- Interpretation: A 1% increase in moonpct is associated with a multiplicative decrease of $\exp(-0.009) = 0.991$ in the estimated mean bat calls.
- We can use `broom::tidy(Bat_GLM, exponentiate = TRUE)` to get the exponentiated results directly.

- If we have an categorical response with more than two levels, we may consider multinomial logistic regression.
- If we have an ordinal response, we may consider ordinal logistic regression.
- For count responses, we may consider zero-inflated poisson or zero-inflated negative binomial models.
- For responses between 0-1, we may consider beta regression.
- For positive continuous responses, we may consider gamma regression.
- Generalized Additive Models (GAM, GAMM) allow splines.

Mixed Models (LMM, GLMM)

- Linear models (LM) and generalized linear models (GLM) assume independent observations.
- When we have repeat observations on the same unit or subject, these observations are correlated.
- Common examples include repeated measures (over time) or blocked designs.
- For both the monarch and bat examples, we have repeat observations at sites.
- If a site has high bat count at one time point, it is reasonable to expect that it would have high bat count at another time point.
- We use mixed models to account for correlated responses.

Mixed Models (LMM, GLMM): Assumptions

We compare the assumptions for LM vs LMM and GLMM.

Assumption	LM	LMM	GLMM
Independence	Yes	No	No
Family	normal/ gaussian	normal/ gaussian	binomial poisson or nb
Variance	Equal	Equal	from family

- The R package lme4 can be used for both linear mixed models (LMM) and some generalized linear mixed models (GLMM).
- The R package glmmTMB has more options for generalized linear mixed models (GLMM).

Monarch and Bat examples

- For both of our example studies, we have repeat observations at sites over time.
- To capture the correlated responses at sites, we use site as a random effect.
- In R, $(1|\text{site})$.
- Everything else can be considered fixed.

Mixed Models: Fixed and Random Effects

- Fixed effects address research questions! These are usually easy for people to identify.
- Random effects are used to capture the study design.

Mixed Models: Fixed and Random Effects

- For this discussion of random effects, I am going to focus on **categorical** predictors because random effects are most often categorical.
- What are we interested in?
 - For fixed effects, we want to compare means for the levels.
 - For random effects, we want to account for variability.
- How were the levels selected?
 - For fixed effects, the levels were of scientific interest.
 - For random effects, the levels are representative of some larger population.
- Consider if someone repeated the study again.
 - For fixed effects, we could run the study again with these levels.
 - For random effects, we would use different levels.

Monarch: Mixed Logistic Regression

```
library(glmTMB)
Monarch_GLMM <- glmTMB(monarchpresence ~ mwcount + year + region
                      + (1|site), family = binomial, data = MonarchData)
summary(Monarch_GLMM)
```

```
Family: binomial (logit)
Formula: monarchpresence ~ mwcount + year + region + (1 | site)
Data: MonarchData
```

AIC	BIC	logLik	deviance	df.resid
2099.3	2127.6	-1044.7	2089.3	2104

Random effects:

Conditional model:

Groups	Name	Variance	Std.Dev.
site	(Intercept)	3.216	1.793

Number of obs: 2109, groups: site, 285

Conditional model:

	Estimate	Std. Error	z	value	Pr(> z)
(Intercept)	399.695596	80.131620	4.988	6.10e-07	***
mwcount	0.016090	0.002375	6.774	1.25e-11	***
year	-0.198711	0.039674	-5.009	5.48e-07	***
regionsouth	-0.268244	0.345904	-0.775	0.438	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Monarch: Analysis and Results Summary

Statistical analysis was conducted with R software and the glmmTMB package (Brooks, 2017). A mixed logistic regression model was fit using monarch presence/absence as the response. Milk weed (*A. Tuberosa*) stem count, year (2015-2024) and region (North or South) were included as fixed effects. Site was included as a random effect to account for repeat observations at sites. Results are presented as odds ratios.

	Odds.Ratio	Coef	SE	p.value
mwcount	1.016	0.016	0.002	0.000
year	0.820	-0.199	0.040	0.000
regionsouth	0.765	-0.268	0.346	0.438

- For count responses, both poisson and negative binomial families are common.
- Negative binomial is a more flexible extension of poisson that allows “over-dispersion”. Negative binomial is common in ecology.
- To decide between poisson and negative binomial, we can fit both models and compare AIC values.
- AIC is very common for model selection in ecology! AIC stands for Akaike Information Criteria.
- The model with the **smaller** AIC is preferred.
- For the Bat data, we find that negative binomial is preferred.

Bat: GLMM with family = poisson

```
library(glmTMB)
Bat_GLMM_Poisson <- glmTMB(batcount ~ brightscore + moonpct + year
                           + (1|site), family = poisson, data = BatData)
summary(Bat_GLMM_Poisson)
```

```
Family: poisson (log)
Formula: batcount ~ brightscore + moonpct + year + (1 | site)
Data: BatData
```

AIC	BIC	logLik	deviance	df.resid
4491.2	4514.8	-2240.6	4481.2	835

Random effects:

Conditional model:

Groups	Name	Variance	Std.Dev.
site	(Intercept)	3.045	1.745

Number of obs: 840, groups: site, 36

Conditional model:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.1396396	0.3334968	-0.419	0.6754
brightscore	0.0735843	0.0725489	1.014	0.3105
moonpct	-0.0030389	0.0007376	-4.120	3.79e-05 ***
year2017	0.1391036	0.0570178	2.440	0.0147 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Bat: GLMM with family = nbinom1

```
Bat_GLMM_NB <- glmmTMB(batcount ~ brightscore + moonpct + year  
                        + (1|site), family = nbinom1, data = BatData)  
summary(Bat_GLMM_NB)
```

```
Family: nbinom1 (log)  
Formula: batcount ~ brightscore + moonpct + year + (1 | site)  
Data: BatData
```

AIC	BIC	logLik	deviance	df.resid
3026.3	3054.7	-1507.2	3014.3	834

Random effects:

Conditional model:

Groups	Name	Variance	Std.Dev.
site	(Intercept)	1.636	1.279

Number of obs: 840, groups: site, 36

Dispersion parameter for nbinom1 family (): 5.11

Conditional model:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.438088	0.278999	1.570	0.1164
brightscore	0.037287	0.053884	0.692	0.4889
moonpct	-0.003591	0.001514	-2.372	0.0177 *
year2017	0.149526	0.112801	1.326	0.1850

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Conclusions

- All models discussed today can include numeric and categorical **predictors**.
- Linear models (LM) have assumptions of independence, normality (of residuals) and equal variance.
- Generalized linear models (GLM) are used for binary **responses** (logistic regression) and count **responses** (Poisson and negative binomial regression).
- Mixed models (LMM, GLMM) allow for correlated responses, for example repeated measures over time on the same plots or subjects.

References and Resources

- Faraway (2016). Extending the Linear Model with R: Generalized Linear, Mixed Effects and Non-parametric Regression Models.
- Cole (2022). EcoCountHelper: an R package and analytical pipeline for the analysis of ecological count data using GLMMs, and a case study of bats in Grand Teton National Park. PeerJ 10:e14509 [<https://doi.org/10.7717/peerj.14509>]
- Prytherch (2022). Shiny app for visualizing logistic regression [https://csu-statistics.shinyapps.io/logistic_regression_data_generator/]

- lme4: linear mixed models and some generalized linear mixed models
- glmmTMB: generalized linear mixed models, with lots of options
- emmeans: helpful when we have categorical predictors
- DHARMA: (interpretable) residual diagnostic plots for GLM, LMM and GLMM
- EcoCountHelper: ecological count data using GLMMs

- [<https://statlab.colostate.edu/>]
- Statistical consulting support available for CSU faculty, staff and students.
- “Drop in” lab: guidance with data formatting, data “wrangling”, data visualization and common statistical analyses.
- Appointments with Stat Lab faculty: Support relating to research proposals, study design, more complicated statistical analyses and manuscript preparation.
- For most projects, there is no charge for general statistical consulting.
- This service is supported by Vice President for Research, Ag Experiment Station, CNS and the Statistics department.

- STAR courses are applied statistics courses for graduate students across CSU (non-Stats students). Note the STAR acronym!
- These courses fill VERY quickly! Early registration is strongly recommended.
- STAR topics courses are offered on a “rotating” basis, but with no set schedule.