Generalised and Mixed Effects Linear Models

A worksheet for Data Science. Written by Hannah Wauchope

21st October 2025

Welcome! Today we're going to be working with a dataset of Hummingbird Counts taken from the Christmas Bird Count. Though this data is freely available online, you do have to make a request to access, which I have done in the past for a separate research project. We should technically request again for this class because its a new use case, so instead I've just changed the years and site names a bit.

In the Christmas Bird Count, volunteers go out once a year and count all the birds they see within a set area. It happens all across North America and has been going over 100 years - so is an incredible data resource. Here, I've taken data on counts of five Hummingbird species across six sites.

Right let's kick off. **Our question for today: "Are hummingbird numbers, in general, increasing or decreasing through time?"**

Load Data and Packages

Run the first few lines of code from the 'LinearModels2_GeneralisedAndMixedEffects.R' to import your data and install packages.

Part 1: Data exploration and prep

Exercise 1 - Visualise

Have an explore of the Humming dataset. How many species do we have? How many sites do we have, and what countries are they in?

We're not going to address it just yet, but take this moment to think about what you've just discovered (and what we discussed earlier in class). How might this data be clustered in terms of independence?

Exercise 2 - Data Prep

We need to modify our Year variable. R interprets it as a continuous variable, and remember that with these it gives intercept and categorical estimates for when continuous variables equal zero. In other words, when you give models Year raw, it will return you intercept estimates at the Birth of Jesus Christ, i.e. Year = 0 (!!)

That's no good, let's scale our years to something more sensible, so that zero represents the first year of our dataset, not 0AD.

```
Humming$YearScale <- Humming$Year - min(Humming$Year)
```

Part 2: Generalised Linear Models

Exercise 3 - Build a GLM

Okay let's get going! We want to model how Hummingbird counts change through time according to year. *But*. We are smart, we know counts don't typically follow a normal distribution. We also know they're discrete (you can't have half a hummingbird), and non-negative (you can't have negative hummingbirds).

What distribution *does* count data tend to follow?

Indeed! So instead of building a Simple Linear Model, which we would almost definitely violate the assumptions of (you can check if you want - code a simple lm correlating YearScale with Count, and plot it. You should see that the first 3 plots are looking funky), we're going to build a *Generalised* linear model.

We're using the function for GLMs from the `lme4` package, and it's nice and easy: `glm` !

This is how it'll look:

```
Mod1 <- glm(Count ~ YearScale, family='WhateverYouWroteAbove_AllLowerCase', data=Humming)
```

Exercise 4 - Understanding GLM output

(Technically we should check model fit before we look at output, but it's important you start learning the following so we'll break the rules for a sec)

Take at at the summary of your model output, specifically the 'YearScale' term of the Coefficients table. *If* this was a normal LM, we would read this as "For every year, Hummingbird Count will go up by _____ on average".

(You should have a very small, decimal number)

That seems odd though, right? We can't have a hummingbird like that.

Remember what I said in the lecture about GLMs, and how they do maths under the hood? Well, model output from a GLM (or GLMM - any model where you're specifying a model family that isn't 'Gaussian' (i.e. normal)) doesn't undo the maths to make it interpretable by us, annoyingly. We have to do that ourselves. In the case of Poisson models, the estimates are currently in log Space, and we have to take the exponential to convert back, as per high school maths.

Try running `exp(...)` where … is the Year Coefficient. Now we no longer interpret this as above.

This is how we interpret Continuous Predictors in a Poisson GLM:

Every year, count increases by a proportion of _____ (where ____ is that number you just got, fill it in).

Another way of thinking about this: when you convert an estimate out of log space, you get a number above or below 1. Above one means a population increase, e.g. 1.03 means the population is increasing by 3% per year. Below one means a population decrease. E.g. 0.97 means the population is decreasing by 3% per year. Ask if that's confusing!

Fill in the blanks of this sentence to finalise your understanding:

According to this GLM, every year, Hummingbird count will (increase/decrease) by ____% on average

(This is a sub par example, as it does look like exp(0.01) has just added 1. Exp is like that for low values, but that's not what its doing - it's taking the exponential. Try running e.g. exp(2) then exp(15) just to prove to yourself its not always that simple!)

Exercise 6 - Check GLM Variance

Okay but really we were getting ahead of ourselves. Let's look at model fit.

How much variance does this model explain? There's no R-squared given in the model output, we have to calculate it ourselves. How rude. Don't worry, it's a simple formula:

Approx R2 = (NullDeviance - ResidualDeviance)/NullDeviance.

Write that formula again, filling in the numbers from what you see in the summary of your model (down the bottom):

Based on that, what percentage of the variance is this model explaining?

Oof. That's a bummer. That's saying that Year only explains ___% of the variance in hummingbird count. In other words, year is absolutely not the main thing determining how many hummingbirds there are.

Perhaps this might not be that surprising…

Exercise 7 - Check GLM Assumptions

SKIP IF YOU'RE FEELING OVERWHELMED (it's more important you get the concepts in the next Exercise, come back to this when you're feeling more confident overall)

As a final check on this model, that we're already pretty suspicious of, let's look at how it meets assumptions. Now, because we've built a GLM we can't use `plot(...)` like before. Until recently we just had to rely on vibes (or being excellent at maths), but thankfully someone built a new R package! We'll take the code from his package `DHARMa`

Rather than `plot(Mod1)`, run:

```
simulateResiduals(fittedModel = Mod1, plot = T)
```

Now, we only get 2 assumptions plots. For the other assumptions, we have to use plots, or our knowledge of the data. Here we have:

- A QQ plot (that, remember, tests for 'normality', or in the case of GLMs that the family we *have* chosen fits well). As with normal QQ plots, the points should follow the diagonal
- A Residuals vs Predicted plot (Which is the same as Residuals vs Fitted, and tests for 'linearity', or that the data behaves how we would expect). We want the three bold lines to follow the three dashed lines, and to be black not red. If they don't, it means there's areas of data that aren't behaving how we expect.

What do you think of this plot?

Exercise 8 - The other variables

Indeed, the assumption plots look *awful*. On the left plot the points are nowhere near the diagonal, and on the right the red lines (which should be evenly distributed horizontally, where the red dashes are) are all over the place. Now what?

Now, sometimes when our assumptions are violated this badly, it because we've missed including key variables that explain the variation in our response variable. Remember, our response variable is 'Count of Hummingbirds'. What are two key variables that might explain how many Hummingbirds there are in each row of the data? (run `head(Humming)` if you're stuck)

1.
2.

Hopefully you've taken all my hints - its Species and Site. Of course. If you walk to a particular place and looked for Hummingbirds, the main thing telling you how many there are isn't "what year is it?".

No, it's the kind of place, and which type of Hummingbird you're looking for. We need to include these variables to set reasonable average estimates for each species and site, and then year can explain whether these averages go up or down through time.

Let's plot this out. Can you make a ggplot of Year vs Count, using geom_point? Add a line (After a + sign) like this: `facet_grid(Species~Site, scales="free")`

Nice!

Part 3: Generalised Linear *Mixed* Models

Exercise 9 - Build a GLMM

Now, remember what we've learnt about random variables. I would like you to build me a model that: estimates average Hummingbird numbers at each site, and tells me whether, on average across all sites, Hummingbirds are going up or down through time. I would like you to account for the fact that Species is likely to also affect variation in Hummingbird abundance, but I don't care about having average estimates for these.

Based on what I just said, fill in the following:

Response Variable:

Categorical Predictor (i.e. Categorical Fixed Effect):

Continuous Predictor (i.e. Continuous Fixed Effect):

Categorical Random Effect:

At least how many groups should a Categorical variable have to be used as a random effect? ____ (in stats language, we use the term 'levels' rather than 'groups')

Are you satisfied your chosen random effect meets this? How many levels does it have? ____

And now write it as model syntax, it should look like this:

```
Response ~ FixedEffect1 + FixedEffect2 + (1|RandomEffect)
```

And we want this to keep being a GLM with a 'poisson' distribution too! (`family = "poisson"`) The function you want to use is `glmer` (yes, it would make more sense if it was 'glmm' - what can you do). Time to code it up - good luck.

Exercise 10 - Check GLMM Assumptions

SKIP IF YOU'RE FEELING OVERWHELMED (it's more important you get the concepts in the next Exercise, come back to this when you're feeling more confident overall)

As per usual - time to check how it fits! Adapt the code from above Exercise 7 and have a look at the assumptions. Remember that in Ecology, we nearly always violate them at least a bit.

I know, they're not perfect. At this stage there is not much more we can do, other than know that these violations (Esp the right graph) mean that our model isn't fitting our data super well. We need to be a little cautious of interpretations - as our model isn't *that* representative of the raw data. Chat to a tutor if you want to understand in more detail why.

Exercise 11 - Check GLMM Variance

SKIP IF YOU'RE FEELING OVERWHELMED (it's more important you get the concepts in the next Exercise, come back to this when you're feeling more confident overall)

The other thing we can check is the variance explained by the model (i.e. the R2). The maths for this is tricky but lucky for us someone developed a function. It's from the 'StatisticalModels' package we loaded. Run `R2GLMER` for your model.

This gives 2 variance estimates - conditional is the whole model, and marginal is just the fixed effects. Fill in the blanks

Conditional: _____(FixedEffect1) and _____(FixedEffect2) and _____ (RandomEffect) explain ____% of the variance in Hummingbird Count

Marginal: _____ (FixedEffect1) and _____ (FixedEffect2) explain ____% of the variance in HummingBird Count

This means that _____ (RandomEffect) explains ___% of the variance (hint - it's Conditional - Marginal!)

Look back at the variance our 'YearScale only' model explained (Exercise 6) - hopefully you can conclude that our model is doing a lot better at explaining what determines how many hummingbirds one counts.

Exercise 12 - Finally Answer the Question

Bravo. Let's look at what our model says! Run: `summary(Mod2)`.

Oo spicy, there's more info here than before. We're not gonna work through all of it in this course. The important thing we'll focus on is our Fixed Effects table.

To check your understanding in the following, first run

`plot(ggpredict(Mod2, c("YearScale", "Site")))` so you have something you can refer to.

Let's practice what we learnt in Exercise 5 first. How does Hummingird Count change each year, on average?

So pretty small increases per year. You can run `confint(Mod2)` to get confidence intervals on this estimate, but remember to take exp(…) from them too!

What about the average number of Hummingbirds per site? Now remember about Categorical variables from last week. We have estimates for each Site, except one. Which one is missing?

But the sneak reveal is that actually the estimate for that site is just the (Intercept). So, what's the average number of Hummingbirds in Arizona A? (We need to run 'exp' here too!)

(Check against the plot to satisfy yourself). What Year, in real terms, is this average for?

What's the average number of Hummingbirds in Guatemala, in that year? (Hint: You must add two numbers together, and you must do this BEFORE you take the exponential!!)

Extension

All done already? Can you figure out how to plot these model prediction lines over the raw data, with one facet per Site-Species combination?

Can you calculate by hand how many Hummingbirds the model predicts in 1984 in Mexico C?

Can you google and figure out how to get the species intercept estimates?