Interpreting Models

Hannah Wauchope

2024-11-07

Welcome! Today we're working with a dataset on Falcons. We'll be consolidating and practicing what we've learnt in previous weeks, and focussing on interpreting model outputs, including interaction terms.

We have a new dataset (Falcons): counts through time of 6 species of falcon at 4 locations.

(Again, this is 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, so instead I've just scrambled 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)

HINTS:

1. When you convert an estimate out of log space (as you need to do for poisson GLM model output), 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!
2. Random effects are used when we need to account for groups in our data, but don't want to get estimates of each group's average.
3. You code interaction terms using a "*" rather than a "+"

WWF Report HINTS:

There are a couple of exercises here that are particularly relevant to the WWF report - I've highlighted them!

Note that we expect you to be thinking critically about your own data for the WWF report. You may not necessarily apply everything as it's taught exactly here. The important thing is that you justify all of your decisions and that they seem sensible. Also, note that if you decide to include extra predictor variables or interaction terms, that if the variables you pick have loads of levels (e.g. if you decide to include 'Country', but your data is across 50 countries) you'll struggle to give a meaningful summary from that - you'll have 50 estimates! You might need to pick a higher level variable (e.g. continent rather than country). Again, this isn't saying you have to do that - it'll depend on your specific species and its data! Make sure to be plotting your data before you make decisions, and thinking about what conclusion you want, and how your data looks.

Our question for today: "What are Falcon numbers doing through time, and across sites?"

Part 1 - Simple GLM

Exercise 1 - Visualise Data

Open the 'LinearModels3' Rscript to work along. Load packages and data.

Now, visualise your data. Make a ggplot that plots Count against Year, and use `facet_grid(Species ~ Site)` to divide this by species and site. Call your plot `PlotFalconData`.

Make a new variable called YearScale that means year starts at 1 (hint: `Falcon$Year – min(Falcon$Year)`), remember from last week - this is because the model estimates intercepts at Year=0 - if we don't scale, we'll get intercept estimates for the year Jesus was Born!

Exercise 2 - Model

Rightio. We're modelling count data, this means we're likely to need at minimum a _____ linear model, with a _____ family specified. (Fill in the blanks!)

Use this info to build a model called `Mod1` that just correlates Count and YearScale (no groups).

Check the residuals and variance explained (it's good practice, though we're not focussing on it today):

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

`r2_mcfadden(Mod1)` (An easier way of getting R2 that I discovered last week! We're all always learning!)

Exercise 3 - Plot Data

Let's plot your model, as this really helps with then interpreting the numbers. Run `plot(ggpredict(Mod1, c("YearScale")))`

Exercise 4 - Intercept Estimates

Now, run `summary(Mod1)`

What is the average number of falcons in Year 0?

Fill in the blanks: we are 95% confident that the average number of falcons in Year 0 falls between ___ and ___.

(remember, you can run `confint(Mod1)` - but then still take the exponential!)

Exercise 5 - Year Slope Estimates

By what average percentage is the modelled population increasing each year? What are the 95% confidence intervals on this estimate?

(run `ggpredict(Mod1, terms=c("YearScale"))` to get the numbers and see if you're looking right - but note that this calculates 95% confidence intervals very slightly differently so don't stress if you're numbers aren't exaaactly identical)

Part 2 - GLMM

Exercise 6 - Accounting for groups

Okay, that's a good basic model. But we know that our data isn't independent, because there are groups. What are the two main groups in *this* dataset? (though get your thinking hats on for next week, there could be variables that we don't have data for!)

Now, in this particular analysis I'm not so interested in the differences between species, I just want an overall average for 'Falcons' (but I do need to account for species in my model). I *am* interested in differences between sites though. I'm also still interested in YearScale

Given all that, that means I would like a _____ linear _____ model. As in the previous exercise, we need to specify a _____ family. We will have two 'fixed effect' predictors, one continuous: _____, and one categorical: _____. We will have one random effect, _____.

Exercise 7 - Build GLMM

Using this info, please build me a model (no interaction terms yet). Write it here first:

Check the residuals, and R2. If you're using lme4 this'll be `R2GLMER(Mod2)` - the 'conditional' is the total variance explained by the model, the 'marginal' is explained by just the fixed effects. If you're using glmmTMB (some people need to cos lme4 was crashing), it's still `r2_mcfadden(Mod2)`. Use the second value, this only gives the 'marginal' variance - i.e. the variance explained by the fixed effects, not the random effects (unfortunately you can't see how much variance is explained by the whole model, we won't penalise for this in the WWF report). How's it looking?

Exercise 8 - Visualise GLMM

Again, let's visualise our model output. Run `plot(ggpredict(Mod2, c("YearScale", "Site")))` (don't worry about the warning if using glmmTMB)

Exercise 8 - Understanding Model Output

Now, run `summary(Mod2)`

Okay, let's remind ourselves of how to interpret the 'Coefficients'. How many Sites do we have in our data, and which Site is not showing? Where do we find the intercept for that hidden site?

Based on that, how many falcons are there on average in Belize1 in Year 0? By what percentage is the population increasing each year?

Remember that all the other categorical estimates are given *relative* to the Intercept. Remember also that categorical estimates should be added together before taking the exponential.

Given that, how many falcons are there in Belize2 in Year0? Is this significantly different from the number of Falcons in Belize1? By what percentage is the population increasing each year? (Hint: I'm not necessarily saying it's any different than that of Belize1…)

Exercise 9 - Check Model Representation of Data (WWF REPORT HINT)

I'm being kind, and have given you some code in the R script (LATER take a minute to understand this plotting code, but in class keep working!). This code takes our model output, and plots it over our raw data to see how well our model is doing at explaining it. Run that code, and write here whether you think the model is doing a good job of representing how falcons are changing through time at each site.

I agree - it's not looking great. The confidence intervals are enormous, and it's showing increases, even though it looks like Mexico 1 is experiencing declines, and Belize 2 maybe no trend at all.

This is where your knowledge of the data is so important in model building. It looks to me like the relationship between YearScale and Count *depends* on which site we're talking about, so if we want to think about sites, we need to tell the model that.

What do we do when the relationship between one predictor and the response depends on another predictor??

(That's right - interaction terms!)

(Important! What if you just wanted to know the overall slope for falcons on average across all sites? In that case you might deliberately *not* interact your variables! But you would lose nuance, but that might be what you're after. Modelling choices all depend both on the data itself, but also what you're interested in. In this exercise, we *do* want to know what's going on for falcons at each site, so we should interact)

Part 3 - Interaction terms

Exercise 10 - Model with Interaction

Can you build me a final model. Make it the same as the one before, but this time, interact YearScale and Site. (You may get a warning on this, don't worry about that). (In a perfect world check assumptions and variance, but skip for now in the interests of time). Write how to do that here first:

Run `plot(ggpredict(Mod3, c("YearScale", "Site")))`. Woah! See how we now have different estimates of the relationship between YearScale and Count for each Site?

Exercise 11 - Interaction Output

Run `summary(Mod3)`. Let's get to work understanding this output. As we see from the graph, the model is now going to estimate not only intercepts for each site, but also YearScale slopes for each site. That's why you can see your three Site estimates, but also three Year:Site estimates (That's the 'YrScle:St' parts, R doesn't like long words so has cut out the vowels). In both cases, note that Belize 1 is missing.

As before, Belize 1's intercept is (Intercept). And now the YearScale estimate for Belize 1 is just 'YearScale'.

Given that, how many Falcons are there in Year 0 in Belize 1? And by what percentage do they change each year?

Check your working against the plot.

Exercise 12 - Interaction Output Pt 2

Now let's get the estimates for Mexico 1. Both numbers are relative - Mexico 1's intercept is relative to Belize 1, and YearScale:Mexico1 is relative to just YearScale.

How many falcons are there in Year 0 in Mexico 1? Is this significantly different from the number in Belize 1?

And by what percentage are falcon populations in Mexico 1 changing per year? Are they going up or down? (Remember: add values together before taking the exponential)

Check against the plot again to check your understanding.

Exercise 13 - Interaction Output Pt 3

To get confidence intervals, run `confint(Mod3)`. Same deal here, e.g. to get confidence intervals for the YearScale of Mexico 1, we would add together YearScale and YearScale:SiteMexico1, and then take the exponential.

Final challenge: Adapt the code from Exercise 4 to plot this new model over the raw data (Hint: you only need to change one thing in that code to adapt it in this case). How does this look?

Wrap up

Reporting Model Outputs (WWF REPORT HINT)

Well done!!! You've made it through, and this is the bulk of complexity we'll learn in this linear model series.

Hopefully you now feel you understand how to build a model, critique it, and understand its output. But how do you communicate your model results in a report?

You should be communicating your output in a number of ways, but especially in something like a report to a charity, you don't want to get too too technical! Your challenge is to communicate your results accurately, but intuitively and clearly.

1. Explain what model you ran, and why, and why you used the predictor variables you did, and why.
2. Explain how well your model meets assumptions and explains variance. This could look like "Though our model did not perfectly meet assumptions, there were no severe violations, so we are reasonably confident our modelled estimates are representative of the data. However, only 5% of variance in the data was explained, meaning there are other variables not included in the model that more strongly affect falcon population numbers."
3. Tell the story of your model output, but focus only on the parts that are most interesting. For example, you might say "According to our model (Table 1, Figure 1), falcon populations in Belize 1 are increasing by an average of 2.7% per year [95% CI: 1.9 - 3.6%]. This estimate is highly significant ($p < 0.0001$), so we are confident in this result. However, falcon populations in Mexico 1 are declining by an average of 4.5% per year [95% CI: declines of 8.1% - 0.8%]. This is significantly different from the estimate for Belize 1 ($p < 0.0001$)."
4. Give a table summarising your model output (more below!)
5. Plot your model output against your raw data, e.g. as per Exercise 4.

Getting Pretty Tables

Say you wanted to display your model output in a table, how might you do that?

We can use the package stargazer! Note that what it DOESN'T do is take the exponential, you still need to do that yourself!

Try running this code:

```
stargazer(Mod3, type="text", style="default", single.row = TRUE)
```

There's loads of things we can play around with in stargazer, for example we might want to make the category labels a bit more clear.

```
stargazer(Mod3, type="text", style="default", single.row = TRUE, ci=TRUE,
  covariate.labels = c("Year", "Belize 2", "Mexico 1", "Mexico 2", "Year*Belize 2",
"Year*Mexico 1", "Year*Mexico 2",    "Intercept"))
```

You could add `ci=TRUE` to get 95% confidence intervals rather than standard errors. `digits=2` would reduce the number of decimal places.

How's it look? Now, change `type="text"` to `type="html"`, and copy the word salad you get out and paste it into an issue or markdown doc on github. Click on the 'preview' tab. What do you see? Nifty right!

Extension

Extension 1 - Random Slopes

Even in our most complex model, we assumed that all species had the same change through time. That's not always reasonable. What can we do? Random slopes! These are essentially interaction terms within the random effect. To say that we think our random factor will interact with year (i.e. species will have different relationships between count and year) we add this `(YearScale|Species)` rather than this `(1|Species)`. Have a go! Plot your data and model - does it fit better?

What do you think?

Extension 2 - Patchy Data

Run `PlotFalconData` again. What do you notice? Do you think our models would fit better, and we'd have better confidence intervals, if we focussed in on only species for which we have a decent amount of data? There are two in particular that seem to have very little data. Have a go at filtering your data and re running models and plotting (paying attention to the fact that if random effects have less than 5 groups they need to become fixed effects), and see what you see.

I'm not necessarily saying you should do this! You may notice our uncertainties are lower. But we've also changed the question we're asking - now we're looking at how *some* falcon species in our data are changing, not all falcon species. You need to use your judgement.