Models for count data

QERM 514 - Homework 8

22 May 2020

Background

This week's homework assignment focuses on fitting and evaluating models for count data. One of your colleagues is interested in the theory of island biogeography and has acquired a data set with which to examine how species richness varies with the area of an island, the island's elevation, the distance to the nearest island, and the area of the nearest island. In particular, her expectation is that the number of plant species should increase with island area, and a plot of the data suggests this to indeed be the case, but her initial modeling effort has yielded the opposite result. Recognizing that she does not have much experience with this type of data analysis, she has turned to you for assistance.

Her data are contained in the accompanying file `plant_richness.csv`, which has the following columns of information:

- `island`: name of the island
- `species`: number of plant species on the island
- `area` the area of the island (km^2)
- `elevation`: the highest elevation of the island (m)
- `distance`: the distance to the nearest island (km)
- `adjacent` the area of the nearest island (km^2)

Questions

a) Plot the number of species versus island area and describe any patterns you observe. Does your colleague's assumption of a positive relationship between richness and area seem to hold?

b) Your colleague explains that she fit the following model, which yielded the surprising result. Fit the model for yourself and verify if there is indeed a negative effect of `area` on `species`. Do the signs of the other coefficients seem to make sense from an ecological perspective? Why or why not?

$$\text{species}_i = \beta_0 + \beta_1 \text{area}_i + \beta_2 \text{elevation}_i + \beta_3 \text{nearest}_i + \beta_4 \text{adjacent}_i + e_i$$

c) Offer one explanation for the unexpected effect of `area` given the apparent relationship in (a). Based on this evaluation, offer a possible suggestion for estimating the effect of `area` on `species`.

d) Does it seem reasonable to use `species` as a response variable in a linear model like the one your colleague fit initially? Why or why not? What would be a more appropriate response variable in a linear model like this?

e) Based upon your knowledge of models for count data, offer a *simple* alternative regression model that models `species` as a function of `area`, `nearest`, and `adjacent`. What are the important components to this model?

f) Fit the model you recommended in (e) and examine the summary information. Does the effect of `area` seem more reasonable in this model? Do you see any problems with this model?

g) Based on your assessment of the model in (f), identify three possible alternatives for estimating the model parameters and their associated uncertainty, and show how you would do so in **R**. How do the these alternative models compare the the estimates in (f).

h) For one of your alternatives in (g), evaluate whether a model that includes only `area` as a predictor is better than a model with all three predictors. Show the **R** code necessary to estimate the model and any test(s) or comparison(s) you might use.

i) Evaluate the diagnostics for your model from (h) with `species` as a function of `area` only. Do you see any problems with this model?