Advancing in R

Module 9: Nonlinear models

Supplementary exercises

In this exercise we will evaluate the effects of altitude (ELEVATION) and plant height (PLANT.HT) on a measure of productivity (TRANSMISSION) using data collected during an MSc field trip in the Cairngorms. You will want to consider the possibility that some of the relationships are curvilinear, and integrate what you learned about the different ways to model nonlinearities with the knowledge you have about multiple regression.

1. Open a new project in a new folder, and start a new script. Save it with a .R suffix to make sure that RStudio interprets commands from it correctly.

2. Read the data from the file "CairngormNL.csv" (in the course materials) into a new object.

3. Check that the file has loaded correctly, and examine the data structure using str().

4. Check the data variables for any noticeable errors. If you see any records that look suspicious, delete the entire row in question. Examine the response and predictor variables, and make a note of any concerns about their distributions.

5. Make some preliminary plots that illustrate the bivariate relationships between the response and the two candidate predictors. Note the sign of any relationships, and flag any nonlinearities that might need to be modelled later. You may also wish to inspect the relationship between the predictor variables because this was not an experiment, and there is potential for the predictors to be multicollinear.

6. Are any transformations useful for normalizing the response and/or linearizing curved relationships?

7. Build a linear model that assesses the influence of PLANT.HT, ELEVATION, and their interaction on TRANSMISSION. Before you examine the model summary, check for variance inflation and examine the diagnostics.

8. Either the diagnostics or the original plot may provide evidence of nonlinearity, so don't rely too much on either one of these alone. If you have any suspicions about nonlinearity, include a second order polynomial for the term in question. Examine the diagnostics of this model and make a note of any concerns.

9. Examine the summary for your minimal adequate model. Try to interpret each of the coefficients. What does the coefficient for an interaction between a second-order polynomial and another predictor represent? How is this different from the term for the interaction between a first order polynomial and another predictor?

10. Use model simplification to find the minimal adequate model for factors affecting TRANSMISSION. Always start by attempting to remove higher order interactions and polynomials before removing any of the main effects.

11. Make some publication quality plots that illustrate the effects of interest. You may want to consider whether showing the partial effects of the variables (e.g., using straight-line added

variable plots) is as evocative of the patterns as showing less representative but perhaps more illustrative plots of curvature (that combine the effects of linear and quadratic coefficients for a single term). It is possible to use {ggplot2} instead of predict() to generate polynomial curves. Can you find help online that directs you to the correct code?

12. Compose several lines of text that would be suitable for a Results section, in which you describe the effects of plant height and altitude on productivity. Because you have conducted model simplification and excluded some coefficients, you will want to explain the justification for this exclusion as well. Include parenthetical statements that cite the appropriate statistical parameters, and refer to your figures as well if you like.

13. Make sure you have annotated your script well enough for your future self or someone else to make sense of it, and then save it.