Advancing in R

Module 4: Multiple Regression

Supplementary exercises

In this exercise we will assess how several predictors [altitude (ELEVATION), soil temperature (SOIL.TEMP.CORR), soil moisture (AVE.MOISTURE), species RICHNESS, and unvegetated area (BARE.GRND)] affected a measure of vegetative productivity (TRANSMISSION) during a MSc field course conducted in 2012 in the Cairngorms.

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 "Cairngorm2012pt.csv" (in your course materials) into a new object. Note that this is not the same file we used in the supplementary exercises for the last practical.
3. Check that the file has loaded correctly, and examine the data structure using str(). Recode any variables that need it.
4. Check the numerical variables for any errors. If you see any records that look suspicious, delete the entire row in question.
5. Examine the distributions of the corrected dataset. Are any of the distributions troubling? Remember that regression makes no assumptions about the distribution of predictor variables. What should you do if no transformations help?
6. Make exploratory plots that illustrate the bivariate relationships between the predictor and each of the response variables. Using these exploratory plots, can you predict the values for the coefficients in a multiple regression? Write down your guesses.
7. What is the null hypothesis for each of the coefficients?
8. Build a linear model that assesses the effect of all predictors on TRANSMISSION (but do not include interactions). Once you have stored the model in a well-named object, check for variance inflation. Are there any values above 10? If yes, what should you do? What about any values above 4? Take whatever steps you deem necessary at this stage, either by proceeding directly to step 9, or by removing problematic variables.
9. Examine the diagnostic plots for your model. Do any of these diagnostic plots give you cause for concern? What should you do about this at this stage?
10. Go through the model simplification procedure prescribed in the lecture, and find the minimal adequate model. Examine its diagnostics, and check again for variance inflation if necessary. In this case the diagnostics will not be brilliant, but since we're not yet prepared for the advanced modelling that would be required to improve this, I suggest we proceed cautiously as if you were happy with this model. Do you agree that this is a reasonable approach?
11. Examine the table of coefficients for your minimal adequate model, and answer the following questions:
    a. What is the sign and magnitude of each significant effect?
    b. Generate the 95% confidence interval for the two coefficients.
    c. How close was your guess about the effect sizes from question 6? Why do you expect to be less accurate with guesses in multiple regressions than in univariate regressions?
12. Which of the significant predictors has the strongest effect on transmission? Can you be confident about this difference in effect size?
13. Prepare publication-quality figures that illustrate your findings. Include one or more legends that explain whether you are illustrating effects from your models or merely illustrating the effects of one variable at a time.
14. Compose a few lines of text that would be suitable for a Results section, in which you describe in general terms your findings, and include the details in parenthetical

statements that cite all the appropriate statistical parameters. Feel free to refer to your figures or a table in this sentence as well.

15. 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.