Advancing in R

Module 3: lm()

Supplementary exercises

In this exercise we will compare how ELEVATION affected the corrected estimates of soil and air temperature obtained during a Stirling MSc field course in the Cairngorms in 2012.

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 "Cairngorm2012.csv" (in the course materials) into a new object. We'll be estimating the effect of altitude (ELEVATION) on the corrected estimates of air (AIR.TEMP.CORR) and soil temperature (SOIL.TEMP.CORR).
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, use R code to delete the entire row in question from a newly named data frame.
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.
6. Make exploratory plots that illustrate the bivariate relationships between the predictor ELEVATION and the two response variables, AIR.TEMP.CORR and SOIL.TEMP.CORR. Using these exploratory plots, can you predict the values for the slopes of the two regressions? Write down your guesses.
7. What is the null hypothesis for each of the two relationships you have just illustrated? You can choose to express the null in terms of the biology, but it is more specific to phrase it in terms of the parameter estimates (i.e., what value does the null take for which coefficient?).
8. Build a linear regression model that assesses the effect of ELEVATION on AIR.TEMP.CORR. Once you have stored the model in a well-named object, examine the diagnostic plots for your model. Do any of these diagnostic plots give you cause for concern? Make annotations that summarize your assessments of the diagnostic plots.
9. For fun, let's compare these diagnostic plots to what we would have seen if we had not properly corrected the data set at the outset. Using the original, uncorrected data frame, build another model. Can you spot the high influence record using the diagnostic plots?
10. Go back to your original, corrected model. Assuming you are satisfied that there are no remarkable deviations from model assumptions, examine the table of coefficients for your model, and answer the following questions:
    a. Is the effect of altitude on air temperature positive or negative?
    b. How strong is the effect, i.e., for every meter of elevation gained, what is the expected change in air temperature?
    c. What is the 95% confidence interval around this expected change?
    d. How close was your guess about the effect size from question 6? If you were within the 95% confidence interval, congratulations! You win a prize in the form of an inflated sense of statistical savvy and self-worth. Well done!
    e. Can you reject your null hypothesis as stated in question 7?
    f. How much of the variation in air temperature is explained by elevation?
11. Repeat steps 8 and 10 for SOIL.TEMP.CORR.
12. Prepare publication-quality figures that illustrate your findings. The approach you use is up to you, but include a line of best fit for each regression and a confidence region around this line, and make sure that your axes are properly labelled. To facilitate comparisons across figures, make sure the y-axis range for both plots is the same.

13. How does the effect of altitude differ for air and soil temperatures? Compose one or two 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. You should refer to your figures in this sentence as well.
14. 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.