Homework 3: Introduction to Statistical Computing Using R

STAT 42100: Modern Statistical Modeling Using R and SAS
STAT 52100: Statistical Computing

Instruction: Collaboration on homework assignments is acceptable, but all write-ups must be done independently and clearly indicate the submitter's understanding of the material. Unclear or disorganized homework may have points removed, even if the content is correct. **A hardcopy of the typed report should be turned in. The pdf file should be submitted electronically on canvas.** An assignment handed in after the deadline is late, and may or may not be accepted. My solutions to the assignment questions will be available when everyone has handed in their assignment.

Edit the program(s) and output together into a single document, showing the lines of code and relevant output produced by SAS/R. Your answers must be easy for the grader to find. A simple structure is, for each part of each question in order, to put these three things:

- Your code
- Your output
- Your answers and explanation

For SAS, the code and output are naturally separate, so this order is good; for R, the code and output can be intermingled, and that is OK. If your assignment is disorganized or otherwise difficult for the grader to deal with, you can expect to lose marks.

You are reminded that work handed in with your name on it must be entirely your own work. It is as if you have signed your name under it. If it was done wholly or partly by someone else, you have committed an academic offence, and you can expect to be asked to explain yourself. The same applies if you allow someone else to copy your work. The graders will be watching out for assignments that look suspiciously similar to each other (or to my solutions). Besides which, if you do not do your own assignments, you will do badly on the exams.

Problem 1: t test Use the data from `fullBumpus.txt` for this problem, and analyze the full dataset together-don't break down by the Group variable.

a) Perform two t-tests to see if the weight of the bird differs by survival status, trying both var.equal=TRUE and var.equal=FALSE. (The latter "adjusts" for unequal variance.) Turn in your two R statements and the corresponding output.

b) With reasonable sample sizes, the t-test is quite robust to (unaffected by) moderate amounts of non-normality. Nevertheless, it is a good idea to check for normality of errors by examining the residuals with a quantile-normal plot. To get residuals for this problem, the easiest method is to re-run the analysis as a simple regression using `res = resid(lm(Weight~Survive, sparrow))`. A nice version of quantile-normal plots with confidence bands is from Brian Junker. To load it, enter source(`"http://math.iupui.edu/~hlwang/STAT521/qqn.R"`). Then create the plot using `qqn(res)`, but don't turn it in. State whether or not you think that the plot shows evidence of sufficient deviation from the reference line to suggest a troublesome degree of non-normality.

c) The t-test is only moderately robust to unequal variance. Unlike the statistical significance of the mean difference, equal vs. unequal variance is easily judged on a side-by-side boxplot. Make a side-by-side boxplot comparing the weight distribution of surviving and perished sparrows. As a rough rule of thumb if the ratio of the IQRs is between 0.5 and 2.0, there is no cause for concern about unequal variances. Roughly what ratio (Survive to Perish, say) do you see?

d) The t-test is non-robust to correlated errors. Correlation is either serial (adjacent subjects are correlated) or by some other grouping, e.g., by nest in this example. The intuition is that, if birds in the same nest are highly correlated in their weights, then there is really not much more information gained by sampling several vs. one bird per nest, but the t-test "thinks" that you have a much larger "n" and therefore inappropriately reduces the estimate of the standard error, resulting in falsely low p-values and falsely narrow confidence intervals. To get a feel for this, load `FakeCor.txt` and make side-by-side boxplots of weight by nest for both WeightA and WeightB (considered as alternate realities). Which one corresponds to correlated (within-nest) errors?

Problem 2: Regression Now we will pretend that the goal of the bird analysis was to model wing length ("Alar") using gender and Weight (without interaction) as explanatory variables.

(a) Turn in the R command to store the `lm()` result in a variable called "mdl". Turn in the result of `summary(mdl)`.

(b) Turn in assignment statements of the form `b0M=, b0F=`, and `b1=` which obtain the estimates of the intercepts and slope from "mdl" using the coef() function. To do this, you need to think about the structural model for the regression, and how it simplifies when "Female" is no longer considered to be a variable, but rather is held constant at 0 (male) or 1 (female). (Do not try to do this by fitting two separate regressions!)

(c) Make a plot with the following layers to summarize the data and model: 1) scatter plot of Alar versus Weight with different color and shape of the points according to whether it's female or not. 2) two linear regression lines of Alar regressed on Weight for males and females. Note that the colors of the points and the colors of the two lines for the males and females need to be consistent.

Also make the residual plot: residual vs. X plot for the model. Put the two residual plots in a single plot and different them in colors and shape by gender. Also draw the horizontal line with residual as 0.

(d) Now repeat the whole process with "mdlI" being the interaction model. You'll need to redefine b0M and b0F, and now introduce b1M and b1F for the separate slopes. Turn in the single plot summarizing the data and the interaction model, with a legend.

(e) Run `anova(mdl,mdlI)` and make a claim whether or not we have good evidence of non-parallel slopes.

(f) Run `confint(mdlI)` and turn in the 95% CI for the difference of slopes (female-male).