Advancing in R

Module 7: predict

Supplementary exercises

In this exercise we will evaluate how population DENSITY affects the ratio of sons and daughters produced by mother wasps, who can control the sex ratio of broods by selectively fertilizing eggs (unfertilized eggs become haploid males). In the course of this exercise, we will discover how to create two-vector responses and try some different approaches for dealing with overdispersion.

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 “sexratio.csv” (in your 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. Examine all the variables, and make a note of any concerns about their distributions. You will notice that there is no variable containing information on sex ratios. Instead, this information is spread across two columns, one providing the number of sons produced at a given density, and the other the number of daughters.
5. For the purpose of visualizing the pattern, create a new variable that is the proportion of offspring that are MALES, i.e., MALES divided by total offspring (MALES + FEMALES). Plot this variable as a function of density. What do you notice about the nature of the relationship? Can you think of any transformation that would linearize this relationship? Because we want to use a binomial model, we shouldn’t transform the response – instead try to come up with a transformation of the predictor that would linearize the relationship. Plot the transformed density variable against the proportion of males to see if you’re right.
6. We could simply build a model of how density (or transformed density) affects this proportion, but if we did that we would be throwing away the extremely valuable information contained in the number of observations of each sex at each density. Instead, you should create a two-vector response variable that includes the raw numbers of observations of each sex, using the cbind command, which combines columns together. Now build a generalized linear model using family=binomial (even though the numbers of individuals are counts, because we’re interested in the proportion assigned to one of two categories, the outcome is binomial) and the two vector response as the left hand side of the formula. For now, let’s try this with the untransformed predictor variable. For example:

   ```
   MOD.1<-glm(cbind(MALES,FEMALES)~DENSITY, data=SEXRATIO,
   family=binomial)
   ```

7. Examine the diagnostic plots for your model using the glm.diag.plots() command from the {boot} library. How terrible do these plots look? Identify the high

influence record. If you like, see whether removing this record improves the model diagnostics (but realize that removing this record amounts to cutting our total sample size by more than half!)

8. Just for fun (even though we know this model is terrible) examine the model summary for your initial model. Is there any evidence of under- or overdispersion? Remember that in order to use the default dispersion parameter of 1, the residual deviance should be between 0.5 and 2 times the residual df.
9. Overdispersion can happen for a number of reasons, including missing predictors or simply a poor model fit. In many cases you may not be able to determine its cause. In such cases, you might have to be satisfied with a lousy model, but then you'll need to use a quasibinomial family designation (or quasipoisson if you are dealing with pure count data), which will adjust the dispersion parameter to reflect your data. Try this now by building a model that is identical to your first one, except that you specify quasibinomial instead of binomial error. Examine the summary for this model, and note the change in the statement about the dispersion parameter below the table of coefficients. Compare the coefficients of this model to those of your initial one. Are they different? What about the standard errors (and consequently the p-values)?
10. We are lucky in that our initial plotting efforts should have provided us with a hint of how to fix the terrible diagnostics from this first model. Perhaps transforming the predictor will improve the model fit and overdispersion. Build a new model of the two-vector response in which the predictor is natural log-transformed density instead of raw density. Examine the diagnostic plots and make a note of any high leverage observations. You will probably find that records with high sample sizes have high influence, but this is to be expected. For today, let's trust the data entry team that all records are correct and proceed as-is.
11. Although it is not necessary for this particular dataset, it's worth mentioning that there are other alternatives to using quasi-families even when the fix is not as "simple" as transforming predictors. Take a moment to search Google for helpful instructions on implementing a few of these. One particularly useful library that you might consider exploring is {aod}, which stands for analysis of overdispersed data. This single package contains functions for negative binomial (which are confusingly designed to deal with overdispersed counts) and betabinomial (for overdispersed binomial data) models.
12. Examine the model summary, and note the dramatic change in residual deviance. Why is this so? Has changing the family altered the ratio of explained and residual variance, or has it merely adjusted model assumptions about that ratio?
13. Can you interpret the model coefficients? Remember that there are two complications in doing so: the predictor variable has been transformed (a natural log transformation), and so has the response (a logistic transformation).
14. Make a publication quality plot that illustrates the effect of density on the proportion of males produced, and illustrates the 95% confidence limits around your fitted relationship. For the purpose of illustration you will probably want to exploit the vector your created for exploring the data earlier, even though this is technically not the variable included in your model. It's probably easiest to plot the curve on the log transformed x-axis, although it is possible to have R display

the raw units on a log scale using the log="x" argument in the initial plot() command.

15. Observe the CI around the line of best fit. Why isn't it symmetrical around the line?
16. Compose a sentence or two of text that would be suitable for a Results section, in which you describe the effect of population density on wasp sex ratio. Include a parenthetical statement that cites the appropriate statistical parameters, and refer to your figure.
17. 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.