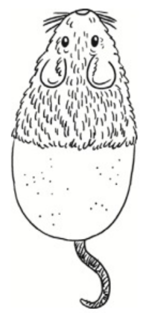
Homework 6

Biology 607

10/9/2014



This week, an amazing paper came out regarding the effect of shaving on daily energy expenditures in voles. `http://jeb.biologists.org/content/early/2014/09/30/jeb.103754.abstract`. It featured one of the best figure 1's that I've seen in a long time. And so, in the spirit of our shaved vole in Figure 1, I'd like us to think about cold rodents.

For this weeks' homework, we'll be working with a data set looking at mice in cold seasonally variable habitats. The data contains four columns: Julian day, temperature (celsius), activity (number of foraging bouts in 6 hours), and food (g food found over 6 hours). We'll investigate this data set and the patterns of correlation in the data in this week's homework.

1 Correlation

We talked about several correlation metrics this week. Pearson's Correlation is when data are drawn from two normal distributions and share a linear relationship. Spearman's correlation is a non-parametric technique when normality is violated that uses correlation of ranks. The Distance Correlation is defined by comparing the pattern of pairwise distances between all values of X and the pattern of pairwise distances between all values of Y. `http://www.imstat.org/aoas/AOAS34INTRO.pdf` provides a lovely intro to it. (Or if you want to get really geeky, see the whole special issue about it at `http://projecteuclid.org/euclid.aoas/1267453929`

1.1 (5 points)

Which `pairs` of variables can we use Pearson's correlation? Which will require non-parametric tests, and if so, which ones? Oh, did `pairs` have funky formatting to it? Might it be a function? Hrm...

1.2 Spearman (5 points)

Spearman's correlation works by calculating correlation based on rank rather than observed value. Write a function to calculate and test Spearman's correlation, and run it on the relationship between temperature and foraging activity. You can compare it to results from `cor.test(method="spearman")`

1.3 Comparing Parametric v. Nonparametric Techniques (5 points)

How do results differ between the three correlation techniques for the relationship between Julian day and temperature? What about activity and food? Use `dcor` from the energy package for the correlation and `dcov.test` to evaluate the null hypothesis for the distance based correlation.

2 Regression

2.1 Diagnostics (5 points)

Aside from eyeballing the relationships, show why we cannot evaluate the relationship between temperature and activity using a ordinary least squares linear regression. What would our tests of regression assumptions tell us about whether linear regression was the right tool?

2.2 Fit (5 points)

Evaluate and discuss the relationship between foraging and food found using linear regression. If it can indeed be evaluated.

3 The Effect of Range

One of the major issues with regression is that estimation and hypotheses testing can be influenced by the spread of your data. Take a look at the relationship between activity and food acquisituion.

3.1 (10 points)

Fit and evaluate the relationship using only activity values between 3 and 9. How does it compare to the fit of the full model? If you answere that the relationship went away, you'd be right - even with 46 data points. Range can matter a great deal - particularly with small sample sizes or high variation. Explore this for yourself. Generate 1000 simulations where you draw out a subset of 10 random rows from the data. Calculate the range of the observed values of activity in each simulation. Also get the p-value for the slope of activity's influence on food. (n.b. `summary(aLM)` produces a list as its output. One item in that list is called coef, which is a matrix from which you can pluck a p-value). Is there some critical range past which the p-value seem to settle down?

4 EXTRA CREDIT: The Effect of Range (10 points)

Man, wouldn't it be cool if you could turn that into a power curve? You can! With something we call a sliding window. Essentially, move across the resulting data, and grab anything with a chosen range + or -, say, 2. Then calculate the power using all p values within that range.