Homework 7

1. Let's simulate how statistical power is affected by sample size and effect size.

Assume a linear relationship between two variables. Because this is a simulation, it doesn't really matter what we call these variables. But to make it more concrete, imagine you are simulating the relationship between net primary production (NPP) and precipitation, for woody terrestrial plant ecosystems. Assume that the intercept of this relationship is 500 (g m^{-2} yr^{-1}), the slope is 0.5, and that primary production is normally distributed around this line with a standard deviation of 500. The predictor (rainfall) is going to vary between 10 and 2000 mm yr^{-1}.

As you'll see, this is a fairly noisy relationship. Let's look at how sample size affects the ability of a linear regression to detect the relationship as significant. Use sample sizes of N = 5, 10, 20, 40, and 80. For each of these sample sizes, do the following simulation 1000 times:

1. Draw N numbers from a uniform distribution between 10 and 2000. This is precipitation.
2. Calculate the expected NPP for these values of precipitation, using the assumed linear relationship.
3. Use the expected NPP to draw random 'observed' values for NPP. These should be normally distributed with standard deviation 500.
4. Fit a linear model for simulated NPP vs. precipitation.
5. Save the estimated slope, and the p-value for the slope.

For each of the sample sizes, record the proportion of p-values (out of 1000 simulations) that are less than 0.05. Finally, plot the statistical power (proportion of significant p-values) vs. the sample size.

There are different ways you could do this, but the easiest is probably to use a nested loop as discussed in lecture.

How much does statistical power to detect this relationship change as the sample size changes?

Also, make a plot of an example simulation for each of the sample sizes, to get sense for what you're simulating.

Now repeat this whole process using a slope of 1 instead of a slope of 0.5. How much does this change in effect size change statistical power at low sample sizes?

2. In lecture 10, I showed some analyses of survey data for Leadbeater's possum. Among other things, I found that the data was zero-inflated. The file 'possum.csv'

contains this data. There are a number of columns, but we're just going to use 'lb', which is the count of possums, and 'stags', which is the number of stags at a site, i.e. the number of hollows that could be used for nesting.

Fit a poisson GLM for possum count vs. stags. Make a scatterplot of the raw data, and plot the fitted curve on top of it. Make a plot of residuals vs. fitted values.

Let's use simulation to get a sense for whether the residual plot looks reasonable. Use the simulate() function to simulate a vector of counts from the fitted model. If you do simulate(your.model), it will generate a dataframe with one column. That column is the simulated data.

Plot the simulated data vs. stags. Fit a poisson GLM for the simulated data vs. stags, and use that to plot the fitted curve. Plot the residuals vs. fitted values.

Repeat this process a total of 4 times. Based on these simulated datasets, and your visual assessment, does it look like the residuals from the real data have a distribution that is consistent with a poisson model?

Visual model assessment is useful, but can't tell you everything. Based on what I did in lecture, we know that the data are zero-inflated. We can use simulation to get a sense for how many 'extra' zeros there are. Use simulate() to generate 1000 simulated vectors of data from the poisson GLM for possums vs. stags. You can do this with simulate(your.model, nsim = 1000). Now you have a dataframe with 1000 columns, and each column is a simulated dataset.

Calculate how many zeros there are in each of the 1000 simulated datasets. The apply() function is your friend here. Make a histogram of #zeros across the 1000 datasets. How does this distribution of #zeros compare to the #zeros in the real data?