Homework #5, Spring 2013 (due April 3rd)

This example is taken from data in Whitlock and Schluter's biostats course. They write:

> "In 1898, Hermon Bumpus collected data on one of the first examples of natural selection directly observed in nature. Immediately following a bad winter storm, 136 English house sparrows were collected and brought indoors. Of these, 72 subsequently recovered, but 64 died. Bumpus made several measurements on all of the birds, and he was able to demonstrate strong natural selection on some of the traits as a result of this storm."

A dataset is posted at `http://phylo.bio.ku.edu/slides/bumpusFromWhitlockSchluter.csv` with the column "survival" indicating whether or not the each bird survived.

For the sake of this homework, we will ignore all of the explanatory variables other than the mass (shown in the "weight(g)" column).

Recall from John's lecture that we can use the logistic link function to model this sort of data (binary response with a continuous explanatory variable). In particular if y_i is the binary outcome of trial i, and x_i is the value of the explanatory variable for that trial, then:

$$\mathbb{P}(y_i = 1) = \frac{1}{1 + e^{-g(x_i)}}$$

where $g(x_i)$ is a linear function:

$$g(x_i) = \alpha + \beta x_i$$

We can treat the event $y_i = 1$ to mean that the bird in trial i survived.

Assignment:

1. The null hypothesis is that mass has no effect on survival. How would you state this hypothesis in terms of constraints on the parameters?
2. Calculate the LRT between the null hypothesis and unconstrained hypothesis. You'll probably want to use the template at `http://phylo.bio.ku.edu/slides/templateParametricBoot.py.txt` to do this (but I cerainly won't penalize you for using some other computational method which allows you to program the likelihood function).
3. Implement parametric bootstrapping.
4. Report:
   (a) best estimates of the parameters for the null and constrained hypotheses,
   (b) the log likelihoods for these hypotheses,
   (c) the approximate P value based on your parametric bootstrapping,
   (d) the critical value for the LRT (based on your parametric bootstrapping)
5. Would it have been acceptable to use the χ^2 distribution according to theory? what about based on your simulations?