Stat 704 Final Exam: Dec. 4th 2014

1. Data analysis (70 points). Burple, Stephens, and Gloopshire (2014) report on a study in the *Journal of Questionable Research* that took place on the Island of Churl in the South Pacific Ocean. Data were collected on the number of minutes Y_i it took $n = 237$ Glippers to learn how to drive a small, motorized car. Glippers are a cat-sized rodent-like mammal with either bright red ($x_{i1} = 0$) or dull green ($x_{i1} = 1$) fur. Two other predictors of interest are the estimated age of the glipper in months x_{i2} (18–25 months in the sample) and the Glippers Maladaptive Score (GMS) x_{i3}, a number from 50 to 100 that summarizes how poor the glipper's vision is (50=mediocre, 100=essentially blind). The data are on the STAT 704 course website near the bottom and are from left to right $x_{i1}, x_{i2}, x_{i3}, Y_i$.

   Find a good, predictive regression model for the time Y_i it takes glippers to learn to drive. Of particular interest to the researchers is quantifying how fur color x_{i1} and age x_{i2} affects the driving time; the GMS variable x_{i3} is included as a concomittant (confounding) variable related to driving time but is not of primary interest.

   Do a thorough, complete job; feel free to use any tools at your disposal, but be sure to show that your final model fits okay. Also be sure to provide careful interpretation of your final model, keeping in mind the goal of the researchers.

2. Short answer (30 points).

   (a) What is the formula for Cook's distance?

   (b) Draw simple linear regression data $\{(x_i, Y_i)\}_{i=1}^n$ where one point has a large Cook's distance but a small residual.

   (c) Draw simple linear regression data $\{(x_i, Y_i)\}_{i=1}^n$ where one point has a large residual but a small Cook's distance.

   (d) Ridge regression helps mitigate what problem in multiple regression? The LASSO accomplishes the same thing but has one additional feature, what is it?

   (e) Weighted least squares is a remedial measure for what?

   (f) Robust (e.g. Huber or median) regression addresses what two problems in multiple regression?