Suggestions for STAT 545 term projects

1) Obtain an appropriate data set and analyze it. There should be enough issues with the data to make it project worthy (as opposed to just a homework problem). For example, you could try different methods and find which is best according to some criterion. The data set may have many predictor variables and devising some way of selecting variables may be challenging. Or you could try different variable selection methods and use some objective criterion to settle on one. One GLM we have not covered yet is log-linear Poisson regression. There are a lot of data sets with counts that can be modeled with this approach, but they often exhibit "over-dispersion" which requires more effort to incorporate in the model. We will talk about the log-linear models beginning in chapter 9.
2) One of the persistent problems in categorical data analysis is the accuracy of the chi-squared approximation to the null distribution when applied to tests involving contingency tables, especially the score test (Pearson's chi-squared test). The original "rule of thumb" I learned was that if *any* cell in the table had expected count less than 5, then the approximation was poor. Later, I heard the rule that if more than 20% of the cells had expected count less than 5 then the approximation was poor. What is the basis for these rules of thumb? This could be a scholarly project where you go looking for sources in the literature for these rules. Maybe you will find other rules of thumb. You can also do simulations under various settings to see how accurate the approximation is for those settings. There is also the question of what to do if the chi-squared approximation isn't valid. For tests of independence or homogeneity, one can use permutation methods to get exact results. For goodness of fit type tests, one can use parametric bootstrap methods to get approximate results, but are these better than the chi-squared approximation of the null distribution?
3) There is a shortage of good methods for ordinal data. We will consider the cumulative logit model for ordinal response variables, which has many shortcomings, but is about the only method I know that is widely used. Simply trying out other proposed methods (there are some in the literature) and comparing them could be a good project. Also, it is common in social sciences to have subjects "rate" items on some ordinal scale such as the "Likert scale" which takes values from 1 to 5. A lot of medical research on pain uses a 1 to 10 scale. These data are often analyzed as if the data are continuous. Using methods like the ACE algorithm discussed in class may provide a more objective way to score the outcomes and apply continuous variable methods. For example, I just read a paper where some political scientists used a Likert scale to assess subjects' attitudes toward a variety of issues, and then applied clustering methods as if the data were simply ordinary continuous variables. They then made many "scientific" conclusions based on the clusters they supposedly found. I am trying to obtain the original data, but there is a lot of potential research to do on such methods.

4) Residuals for binary outcome models are very problematic and don't seem to be helpful. There are a lot of potential projects here. I have posted a project from last year that looked at residuals in simulated data, and the residuals from fitting the correct model looked weirder than for an incorrect model, which is not unusual. Some research into what info there is in the literature on residuals for binary outcomes combined with some simulations or analysis of real data sets could make a good project. I have some ideas for alternatives to existing methods for computing residuals as well.
5) The proportional hazards (Cox regression) model is known to be a GLM. A project explaining how this model fits into the GLM framework could work. It is not enough to just show this, but one could maybe look at other GLM type models, or show how GLM machinery can be used in proportional hazards modeling.
6) We have only mentioned nonparametric GLMs without going into too much detail. Implementing a nonparametric GLM is relatively easy using the optional variable "weights" in the glm function of R. For example, to get a nonparametric estimated value at a given x, input kernel weights of the form $w_i(x) = K((x-x_i)/h)$ where K is a kernel function and h is the bandwidth parameter.
7) One way to test goodness-of-fit for binary response GLMs is to fit a non-parametric estimate to the probability function and then look at the sum of squared differences between the parametric and nonparametric fits. One can get a null distribution for the test of the null hypothesis that the true model is the parametric model by using parametric bootstrap methods (simulate new data sets from the fitted parametric model).