Optional Homework
Stat 431, Summer 2012

Due Thursday Aug 9, 2012 before class begins

All the questions are bonus questions. Do as much as you can or do nothing. I would suggest you work on the project first and whenever you have time, start doing some of these questions below.

1. This question pertains to analyzing a very famous data set, collected by none other than R.A. Fisher (aka the father of statistics). In fact, it's so famous that all R installations carry this data set. Type in iris and you'll get a data set R.A. Fisher collected. The data set is a collection of iris flowers and their various physical characteristics.

   (a) We want to build a model that differentiates between setosa species from the rest. Construct a logistic regression model where $Y_i = 1$ indicates that the iris flower is seposa and $Y_i = 0$ indicates that the iris flower is not seposa and all the X's are your independent variables. What is the prediction error?

   Hint: Prediction error is defined to be whether the model correctly predicts $Y_i = 1$ or $Y_i = 0$. Specifically, it is

   $$\frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

   To obtain $\hat{Y}_i$, simply use the

   `predict(model,type=''response'')`

   function in R. For simplicity, designate probabilities that are greater than 0.5 to be $\hat{Y}_i = 1$ and those that are less than 0.5 to be $\hat{Y}_i = 0$

   (b) Are the X's useful in determining whether the iris flower is seposa? Conduct a hypothesis test, set up the hypothesis, define the reduced and the full models, and compute the p-value. Is the null rejected at $\alpha = 0.05$?

   (c) We want to know whether sepal features matter, controlling for the petal features, in predicting whether the species is seposa or not. Conduct a hypothesis test, set up the hypothesis, define the reduced and the full models, and compute the p-value. Is the null rejected at $\alpha = 0.05$?

   (d) We want to know whether model only with sepal features as independent variables is better than model only with petal features as independent variables. Rank which model is better by using the three information criterion. For each information criterion, which model seems to be the best?

2. Here, you will predict Facebook's stock price (or at least attempt to in the short run...)

(a) Plot the closing price of Facebook's stock as a function of time. What kind of trend do you see?
Hint: The first point in the data is the last collected time point. So you need to revert the ordering of the data set so you start with the earliest day on the left side of the x-axis

(b) Build a time series model that includes a trend term and a lag term. Choose an appropriate trend term and an appropriate number of lag terms to include by using the test shown in the time series lecture. Plot the predicted price with the actual price and state the final model.

(c) Test whether the trend term is important, given the lag term. State the hypothesis, the test statistic, and compute the p-value.

(d) What is the closing value of Facebook's stock next Friday (August 10, 2012)?

3. Consider the simple logistic regression where

$$\log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X \tag{1}$$

Show that the deviance between the model with only the intercept and equation (1),

$$\Delta = -2\log(lk_{red}/lk_{full})$$

has a Chi-square distribution as n goes to infinity under the null hypothesis that $\beta_1 = 0$. Here, lk_{red} represents the likelihood for the reduced model and lk_{full} represents the likelihood for the full model.
Hint: Solving this question will automatically get you an A. This is mathematically challenging and requires Taylor expansions and moment generating functions