Stats for Clinical Trials, Math 150
Jo Hardin
Homework # 8, Due: Thurs, 11/4/10

Using the Duchenne Muscular Dystrophy (DMD) data:

- table_no: The table number in a book of tables, in this case table number 38
- id: Measurement identification number (1-209)
- obs: Number of measurements of taken from a given patient
- hosp_id: Patient identification number
- age: Patients age (in years)
- month: The month that a given serum sample was drawn (1-12)
- year: The year that a given serum sample was drawn (19XX)
- CK: Creatine Kinase activity (international units/mL)
- H: Hemopexin concentration (mg protein/100 mL, mg%)
- PK: Pyruvate Kinase activity (international units/mL)
- LD: Lactate Dehydrogenase activity (international units/mL)
- group: The carrier status of women ("carrier" or "normal")

Percy et al. sought to determine whether an assessment of both CK activity and H levels would provide a better indication of carrier status. The study found that H levels alone distinguish only 27% of carriers, but in combination with CK activity distinguish 83% of carriers, allowing the researchers to conclude that the combination of assays is more effective in determining carrier status. (Percy ME, Andrews DF, and Thompson MW (1980). *Duchenne Muscular Dystrophy Carrier Detection Using Logistic Discrimination: Serum Creatine Kinase and Hemopexin in Combination.* Am. J. Med. Gen. 8:3 97-409.)

1. Make a scatterplot of H vs. loc(CK); use one plotting symbol to represent the controls on the plot and another to represent the carriers. Does it appear from the plot that these enzymes might be useful predictors of whether a woman is a carrier? Explain.
```
group.symbol <- ifelse(group=="normal", 19, 12)
plot(H, log(CK), pch=group.symbol)
```

2. Fit the logistic regression of carrier on CK and CK-squared (`CK^2`). Does the CK-squared term significantly differ from 0? Next fit the logistic regression of carrier on log(CK) and `(log(CK))^2`. Does the squared term significantly differ from 0? Which scale (untransformed or log-transformed) seems more appropriate for CK?
   Hint1: square the term outside of the logistic regression model `CK2 <- CK^2`
   Hint2: just add the term like you would add any other explanatory variable
   `glm(group ~ CK + CK2, family="binomial")`

3. Fit the logistic regression of carrier on log(CK) and H. Report the coefficients and standard errors.

4. Carry out a likelihood ratio test (also called a drop-in deviance test) for the hypothesis that neither log(CK) nor H are useful predictors of whether a woman is a carrier.

5. Starting with all main effects and all interactions, use the `drop1` command to go backward to a stable model that you would report (use only the age, CK, H, PK, and LD variables (no squares or other transformations)). Report the final model.

6. Repeat the previous problem starting with nothing and using the `add1` command to go forward to a stable model. Report the final model.

7. Comment on any similarities and differences you see in the backward and forward models you created.

Note: I don't know how R will code the response (currently labeled with a string variables: normal/carrier). In the work above, you need to make sure you know which is which. You might start by redefining a different variable to use so that you know what R is modeling as the odds. That is, R **always** models the odds of getting a success, i.e. a "1".

```
new.group <- ifelse(group=="normal", 0 , 1)
```