APPLIED STATISTICS
TUTORIAL 9

Question 1 in Tutorial 8 (Con'd, ex from Chapter 20 of the class text "The Statistical Sleuth")

Duchenne Muscular Dystrophy (DMD) is a genetically transmitted disease, passed from a mother to her children. Boys with the disease usually die at a young age; but affected girls who usually do not suffer symptoms, may unknowingly carry the disease, and may pass it to their offspring. It is believed that 1 in 3300 women are DMD carries. A woman might suspect she is a carrier when a related male child develops the disease. Doctors must rely on some kind of test to detect the presence of the disease. The file "DMD.csv" contains levels of two enzymes in the blood, creatine kinase (CK) and hemopexin (H) for 38 known DMD carries and 82 women who are not carriers (controls). It is desired to use these data to obtain an equation for indicating whether a woman is a likely carrier.

b) Fit the logistic regression of carrier on CK and CK^2. Does the CK^2 term differ significantly from 0? Next fit the logistic regression of carrier on log(CK) and $[\log(CK)]^2$. Does the squared term differ significantly from zero? Which scale (untransformed or transformed) seems more appropriate for CK?
d) Carry out a drop-in-deviance test for the hypothesis that neither log(CK) or H are useful predictors of whether a woman is a carrier.

Question 1 (ex from Chapter 20 of the class text)

The file "shuttle.csv" contains data on the launch temperatures and an indicator for O-ring failure for 24 space shuttle launches prior to the space shuttle Challenger disaster of January 27, 1986.

Fit the logistic regression of Failure (code failure as "0") on Temperature (include temperature as the only term in the fitted model). Now fit a second logistic regression of Failure (code failure as "1") on Temperature. Reconcile the coefficient estimates from the two models, that is, clearly show that they will lead to the same conclusions about the relationship between Temperature and Failure.

Question 2

To investigate the outbreak of a disease spread by rodents, people were sampled from two different locations in a particular town. The sampled individuals were tested to see whether they were carrying the disease. The response variable is an indicator variable the disease being present (1 if present, 0 otherwise). The available explanatory variables are socioeconomic status (categorical variables three categories) (X_2, X_3), age (X_1), and the location that the person was sampled from (categorical with two categories). The output from a logistic regression model fitted to this data is provided below.

```
Call: glm(formula = Y ~ X1 + X2 + X3 + X4, family = binomial(link = logit))
```

```
Deviance Residuals:
       Min         1Q     Median         3Q     Max
 -1.655179 -0.7529131 -0.4787575 0.8558047 2.09767

Coefficients:
                  Value Std. Error    t value
(Intercept) -2.31293438 0.64244340 -3.6002150
         X1  0.02975008 0.01350009  2.2036958
         X2  0.40879015 0.59894404  0.6825181
         X3 -0.30525427 0.60400966 -0.5053798
         X4  1.57474897 0.50154323  3.1398071

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 122.3176 on 97 degrees of freedom

Residual Deviance: 101.0542 on 93 degrees of freedom

Number of Fisher Scoring Iterations: 4

Correlation of Coefficients:
   (Intercept)         X1         X2         X3
X1 -0.6585092
X2 -0.4769266   0.1421288
X3 -0.5178919   0.0898341  0.4096387
X4 -0.5063180   0.0497322  0.0428483  0.2055461
```

a) Write down the fitted logistic regression model. What is the estimated increase in the log-odds of the disease being present for each 5-year increment in age, everything else held constant? Provide a 95% confidence interval for your estimate.

b) Amongst a group of 20 individuals all aged 30 years and all with $X_2 = 1, X_3 = 0$ and $X_4 = 0$, how many would you expect to have the disease present?

c) You have found another three observations that were not included in your dataset when the model above was fitted. These three observations are provided in the table below. Based on these three values, does the cut-off probability 0.5 result in the three observations being correctly classified (predicted)? You must show working.

X_1	X_2	X_3	X_4	Y
44	0	1	1	1
11	0	1	1	0
3	1	0	1	1