**STAT 425/525 - Homework 7**
**Due Monday October 31, 2016**

*Unless otherwise indicated you can use computer software to do the following problems.*

1. The analyses presented below are from a study about the relationship between waking with a sore throat after surgery as a function of duration (in minutes) and type of device. We have

$$Y = \begin{cases} 1 & \text{sore throat} \\ 0 & \text{no sore throat} \end{cases} \qquad type = \begin{cases} 1 & \text{mask airway} \\ 0 & \text{tracheal tube} \end{cases}$$

   The data will be provided to you in the file `sorethroat.csv`. R code that may prove useful is also attached.

   (a) Fit an additive model to the data and give me the summary results.

```
sore.fit<-glm(y~type+duration,family=binomial,data=sore.data)
summary(sore.fit)
```

   On the logit scale write down the two equations corresponding to mask and tube, respectively.

   (b) Estimate the ratio of the odds of a sore throat when using a mask airway to a sore throat when using the tracheal tube. Provide me with an approximate 95% confidence interval for the true odds ratio. Interpret the interval in terms of the problem.

   (c) Estimate the ratio of the odds of a sore throat for a surgery of 40 minutes to the odds for a surgery of 30 minutes. Provide me with an approximate 95% confidence interval for the true odds. Interpret the interval in terms of the problem.

   (d) Fit a model with an interaction between type and duration. Summarize the results and write down (on the logit scale) the two equations corresponding to mask and tube, respectively.

```
sore.fit2<-glm(y~type*duration,family=binomial,data=sore.data)
summary(sore.fit2)
```

   (e) Using the interaction model estimate the ratio of the odds of a sore throat for a surgery lasting 30 minutes when the mask is used to the odds of a sore throat surgery lasting 30 minutes when a tracheal tube is used. Provide me with an approximate 95% confidence interval. Interpret the interval in terms of the problem.

   (f) Do we need the interaction term? Carry out both a Wald test and a likelihood ratio test. Summarize the results and draw a conclusion in terms of the problem.

2. Kyphosis is a disfiguring forward flexion of the spine following spinal surgery. Age in months of 18 subjects with kyphosis ($y = 1$) and 22 subjects without kyphosis ($y = 0$) are given below.

```
y<-c(rep(1,18),rep(0,22))
age<-c(12,15,42,52,59,73,82,91,96,105,114,120,121,128,130,139,139,157,
1,1,2,8,11,18,22,31,37,61,72,81,97,112,118,127,131,140,151,159,177,206)
```

(a) Fit two logistic regression models with age as the predictor. One model only incorporates age as a linear term while the second has age and age squared in it.

```
fit<-glm(y~age,family=binomial)
fit.q<-glm(y~age +I(age^2),family=binomial)
```

Compare the two models. Does it appear that the relationship is quadratic? Justify your answer.

(b) Interpretation of quadratic effects can be problematic. A graphical assessment is often helpful. Plot the estimated odds of kyphosis versus age. Provide me with a copy of the plot.

```
plot(age,fitted(fit.q),xlab="age",ylab="fitted odds")
```

(c) Summarize the relationship.

(d) STAT 525 Only: At what age is the estimated odds of kyphosis the greatest (I want a general answer given as a function of the estimated regression coefficients)? Give an approximation for the variance of this estimated age. Now use the general formulas you just derived to give an estimate of the age and the variance of that estimate for the kyphosis problem.

3. Duchenne Muscular Dystrophy (DMD) is a genetic disease transmitted from mothers to their children. Male offspring generally do not live very long but females may be silent carriers - they do not get sick but are capable of transmitting DMD to their offspring. Blood levels of 2 enzymes (creatine kinase (CK) and hemopexin (H)) were evaluated as possible screening tools. The data are available in an R package `Sleuth3`.

(a) Fit a model with the explanatory variables `log(ck)` and `h`. Summarize the results.

```
require(Sleuth3)
MD.dat<-ex2012
names(MD.dat)<-c("group","ck","h")
attach(MD.dat)
group<-ifelse(group=="Control",0,1)
fit<-glm(group~log(ck)+h,family=binomial)
summary(fit)
```

(b) Create a classification table using a cutpoint of c = 38/120. Compare the accuracy, sensitivity, speci

city, $PV+$, and $PV-$ values (summarize your results in a table). Interpret these measures in terms of the problem. Be careful with $PV+$ and $PV-$ as these are case-control data and those values are not applicable for a general population.

```
pi.hat<-predict(fit,type="response")
y.hat<-ifelse(pi.hat>=38/120,1,0)
table(group,y.hat)
```

(c) Install the Epi package if you have not already done so and use the ROC function to create a ROC curve using fitted results. Show the plot in your write-up.

```
require(Epi)
# additive model
ROC(fitted(fit.1),group,plot="ROC")
```

Compare the results commenting on

i. the optimum cutpoint

ii. Sensitivity, Specificity, $PV+$, and $PV-$. Recall that the $PV+$ and $PV-$ values in the plot are incorrect but you can use them to get the correct values.

iii. AUC and the implication it has in terms of using these two enzymes as a screening tool.