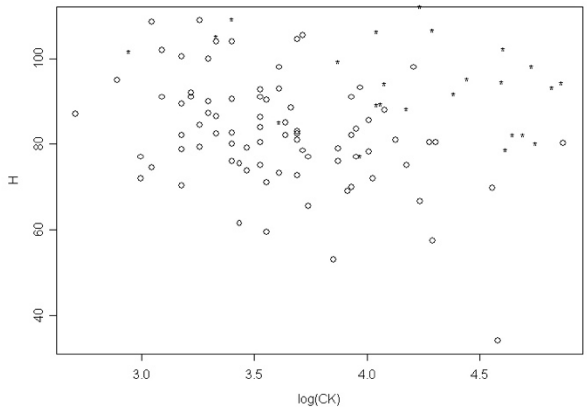
APPLIED STATISTICS
TUTORIAL 9 SOLUTIONS

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

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 that1 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. It is desired to use these data to obtain an equation for indicating whether a woman is a likely carrier.

a) Make a scatterplot of H versus log(CK); use one plotting symbol to represent the controls on the plot and another to represent the carriers. Does it appear that these enzymes might be useful predictors of whether a woman is a carrier?

```
DMD=read.table("DMD.csv",header=T,sep=",")
names(DMD)
CK=DMD$CK
H=DMD$H
group=DMD$GROUP
plot(log(CK[group==group[1]]),H[group==group[1]],xlab="log(CK)",ylab="H")
points(log(CK[group==group[83]]),H[group==group[83]],pch="*")
```



The "*" represent the carriers. The points for the carriers and controls show a clear separation. The two variables should provide a good way of distinguishing between the two groups.

b) Fit the logistic regression of carrier on CK and CK-squared. Does the CK-squared 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?

```
DMD.logit1=glm(group~CK+I(CK^2),family=binomial(link=logit))
summary(DMD.logit1)


Call:
glm(formula = group ~ CK + I(CK^2), family = binomial(link = logit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.50614  -0.03892   0.37943   0.51824   2.27518

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.181e+00  7.272e-01   5.749 8.96e-09 ***
CK          -5.805e-02  1.301e-02  -4.460 8.18e-06 ***
I(CK^2)      5.060e-05  3.286e-05   1.540    0.124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149.840  on 119  degrees of freedom
Residual deviance:  85.435  on 117  degrees of freedom
AIC: 91.435

Number of Fisher Scoring iterations: 9
```

The fitted logistic regression is:

$$\text{logit}(\hat{\pi})=4.18-0.058CK+0.00005CK^2$$

The squared term is not significant. The test statistic is 1.54 with a corresponding p-value of 0.124.

Using the transformed variables

```
logCKsqr=(log(CK))^2
DMD.logit2=glm(group~log(CK)+logCKsqr,family=binomial(link=logit))
summary(DMD.logit2)


Call:
glm(formula = group ~ log(CK) + logCKsqr, family = binomial(link = logit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.39251  -0.03075   0.38037   0.50190   2.28852

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.830     16.309  -0.603    0.547
log(CK)        8.568      8.366   1.024    0.306
logCKsqr      -1.453      1.064  -1.365    0.172

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149.84  on 119  degrees of freedom
Residual deviance:  84.98  on 117  degrees of freedom
AIC: 90.98

Number of Fisher Scoring iterations: 7
```

In this case the squared term is not significant. The test statistic is 1.4 with a corresponding p-value of more than 5%.

The model with the transformed variables is preferred because of its smaller deviance (both models have same number of explanatory variables).

c) Fit the logistic regression of carrier on log(CK) and H.

```
DMD.logit3=glm(group~log(CK)+H,family=binomial(link=logit))
summary(DMD.logit3)


Call:
glm(formula = group ~ log(CK) + H, family = binomial(link = logit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.60372  -0.09903   0.16697   0.38782   1.89707

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 28.91300    5.80030   4.985 6.20e-07 ***
log(CK)     -4.02041    0.82909  -4.849 1.24e-06 ***
H           -0.13652    0.03654  -3.736 0.000187 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149.840  on 119  degrees of freedom
Residual deviance:  61.992  on 117  degrees of freedom
AIC: 67.992

Number of Fisher Scoring iterations: 7
```

The fitted logistic regression is:

$$\text{logit}(\hat{\pi})=28.9-4.02\log(CK)-0.14H$$

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.

```
> DMD.logit3R=glm(group~1,family=binomial(link=logit))
> anova(DMD.logit3R,DMD.logit3,test='Chisq')
Analysis of Deviance Table

Model 1: group ~ 1
Model 2: group ~ log(CK) + H
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       119    149.840
2       117     61.992  2   87.848 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This means the p-value is approximately zero. The data suggests that at least one of the variables is important.

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.

```
> space=read.table("shuttle.csv",header=T,sep=",")
> temp=space$TEMP
> failure=space$FAILURE
>
> indFAIL<-ifelse(failure=="Yes",1,0)
> indSucc<-ifelse(failure=="No",1,0)
>
> space.logit1=glm(indFAIL~temp,family=binomial(link=logit))
> space.logit2=glm(indSucc~temp,family=binomial(link=logit))
>
> summary(space.logit1)

Call:
glm(formula = indFAIL ~ temp, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2125  -0.8253  -0.4706   0.5907   2.0512

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.87535    5.70291   1.907   0.0565 .
temp        -0.17132    0.08344  -2.053   0.0400 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 28.975  on 23  degrees of freedom
Residual deviance: 23.030  on 22  degrees of freedom
AIC: 27.03

Number of Fisher Scoring iterations: 4

> summary(space.logit2)

Call:
glm(formula = indSucc ~ temp, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0512  -0.5907   0.4706   0.8253   1.2125

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.87535    5.70291  -1.907   0.0565 .
temp          0.17132    0.08344   2.053   0.0400 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.975  on 23  degrees of freedom
Residual deviance: 23.030  on 22  degrees of freedom
AIC: 27.03

Number of Fisher Scoring iterations: 4
```

From this output we can see that the estimates of the coefficients have their signs switched. We know that:

$$\log it(\pi) = \beta_0 + \beta_1 X$$

$$\Rightarrow \pi = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

$$1 - \pi = 1 - \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

$$\Rightarrow 1 - \pi = \frac{\exp(-\beta_0 - \beta_1 X)}{1 + \exp(-\beta_0 - \beta_1 X)}$$

But it is $1-\pi$ that we are modelling when we change the coding of our response.

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.

$$\text{logit}(\hat{\pi}) = -2.31 + 0.0297X_1 + ... + 1.57X_4$$

$$5 \times 0.0297 \pm 2 \times 5 \times 0.0135 = (0.0135, 0.2835)$$

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?

$$\text{logit}(\hat{\pi}) = -2.31 + 0.0297 \times 30 + 0.408$$

$$\hat{\pi} = \frac{\exp(-2.31 + 0.0297 \times 30 + 0.408)}{1 + \exp(-2.31 + 0.0297 \times 30 + 0.408)} = 0.267$$

$$20 \times 0.267 = 5.3$$

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

Probabilities from fitted model are 0.57, 0.33, and 0.44, respectively. So the answer is no.