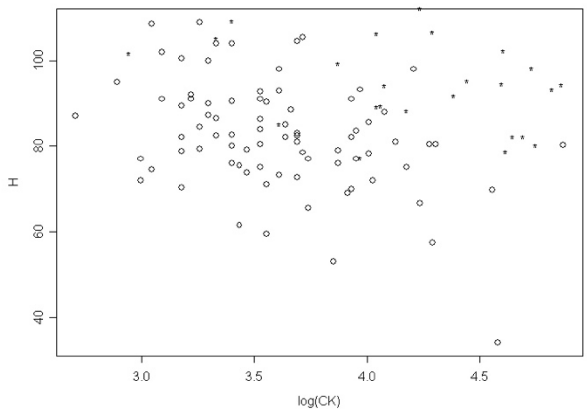
APPLIED STATISTICS
TUTORIAL 8 SOLUTIONS

Question 1 (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. Next fit the logistic regression of carrier on log(CK) and $[\log(CK)]^2$.

```
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:

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

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
```

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:

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

e) Typical values of CK and H are 80 and 85. Suppose that a suspected carrier has values of 300 and 100. What are the odds she is a carrier relative to the odds that a woman with typical values (80 and 85) is a carrier?

```
> CKtypical<-80
> CKsuspected<-300
> Htypical<-85
> Hsuspected<-100
>                                                                    ODDS.typical<-
exp(DMD.logit3$coef[1]+DMD.logit3$coef[2]*log(CKtypical)+DMD.logit3$coef[3]*Htypical)
> ODDS.typical
(Intercept)
  0.7346736
>                                                                  ODDS.suspected<-
exp(DMD.logit3$coef[1]+DMD.logit3$coef[2]*log(CKsuspected)+DMD.logit3$coef[3]*Hsuspect
ed)
> ODDS.suspected
(Intercept)
0.000466593
> 1/(ODDS.suspected/ODDS.typical)
(Intercept)
   1574.549
```

The fitted logistic model was:

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

$$\Rightarrow \frac{\hat{\pi}}{1-\hat{\pi}} = \exp(28.9-4.02\times\log(CK)-0.14\times H)$$

$$\frac{\exp(28.9-4.02\times\log(300)-0.14\times 100)}{\exp(28.9-4.02\times\log(80)-0.14\times 85)}$$ =ODDS.suspected/ODDS.typical

So the odds that the suspected woman is a carrier are `1/(ODDS.suspected/ODDS.typical)= 1574.549` times the odds that a woman with typical values is a carrier. Note we are modelling a women **not being a carrier** as 1 in R **by default** (since we did not use the "ifelse" function but used the categorical "group" directly as the response in the "glm" function), so we need to invert the value of `ODDS.suspected/ODDS.typical`.