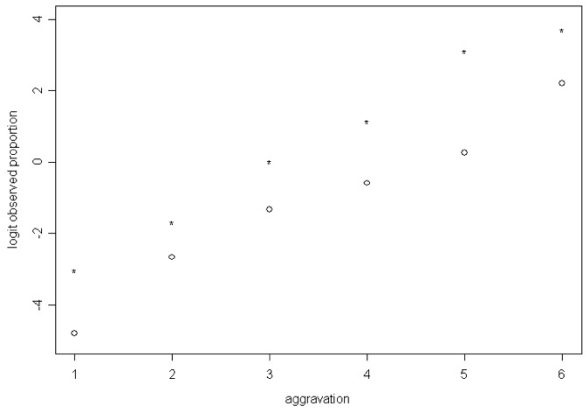
APPLIED STATISTICS
TUTORIAL 10 SOLUTIONS

Question 1 (ex from Chapter 21 of the class text)

The data "penalty.csv" contains data involving death penalty and the race of the victim. The data contains category of the murder or aggravation (higher numbers correspond to more viscous murders), the race of the victim, the number receiving death, and the number not receiving death. The response variable is the number who receive the death penalty.

a) Plot the logit of the observed proportions versus the level of the aggravation. Use different plotting symbols for white and black victims.

```
penalty=read.table("penalty.csv",header=T,sep=",")
names(penalty)
aggravation=penalty$AGGRAVATION
victim=penalty$VICTIM
death=penalty$DEATH
nodeath=penalty$NODEATH
indblack=ifelse(victim==victim[2],1,0)
logitprop=log((death+0.5)/(nodeath+0.5)) #empirical logit. Fixes up problems of
   undefined logit
plot(aggravation[indblack==1],logitprop[indblack==1],ylab="logit observed
   proportion",xlab="aggravation",ylim=c(-5,4))
points(aggravation[indblack==0],logitprop[indblack==0],pch="*")
```



Note the use of the empirical logit in the above code. This plot seems to indicate that people who murder white victims are more likely to receive the death penalty.

b) Fit the logistic regression of death sentence proportions on aggravation level and an indicator variable for race of victim.

```
Y=cbind(death,nodeath)
death.logit=glm(Y~aggravation+indblack,family=binomial(link=logit))
summary(death.logit)
Call:
glm(formula = Y ~ aggravation + indblack, family = binomial(link = logit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.93570  -0.22548   0.05142   0.65620   1.01444

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.8653     0.6004  -8.103 5.37e-16 ***
aggravation   1.5397     0.1867   8.246  < 2e-16 ***
indblack     -1.8106     0.5361  -3.377 0.000732 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212.2838  on 11  degrees of freedom
Residual deviance:   3.8816  on  9  degrees of freedom
AIC: 31.747

Number of Fisher Scoring iterations: 4
```

The fitted logistic regression is:

$$\text{logit}(\hat{\pi})=-4.87+1.54\text{aggravation}-1.81\text{black}$$

c) Report the p-value from the deviance goodness-of-fit test for this fit.

From the summary output we find that the deviance statistic is 3.88 with 9 d.f. The corresponding p-value is 0.9191.

```
> 1 - pchisq(3.88, 9)
[1] 0.9191315
```

d) Test whether the coefficient of the indicator variable for race is equal to 0.

From the summary() output we see that the test-statistic for this hypothesis is -3.4 with a corresponding p-value of 0.0007. The data suggests that race is an important variable. The odds of death are lower if the victim was black.

e) Construct a 95% CI for the same coefficient, and interpret it in a sentence about the odds of death sentence for black-victim murderers relative to white-victim murderers, accounting for aggravation level of the crime.

A 95% CI is (-2.86, -0.76). If we exponentiate this CI we get (0.057,0.47). This says that they odds of getting the death penalty for murdering a black victim are estimated to be between 0.057 to 0.47 times the odds for murdering a white victim, accounting for the aggravation level of the crime.

Question 2 (ex from Chapter 21 of the class text)

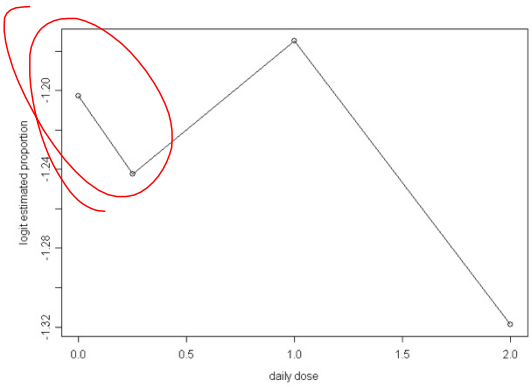
Between December 1972 and February 1973, a large number of volunteers participated in a randomized experiment to assess the effects of large doses of vitamin C on the incidence of colds. The subjects were given tablets to take daily, but neither subjects nor the doctors who evaluated them were aware of the dose of vitamin C contained in the tablets. In the display below are the proportions of subjects in each of the four dose categories who did not report any illnesses during the study period.

$\log\left(\frac{0.231}{1-0.231}\right) = \log\left(\frac{267}{1158-267}\right)$

Display 21.18 Vitamin C and colds

Daily dose of vitamin C (g)	Number of subjects	Number with no illnesses	Proportion with no illnesses
0	1,158	267	0.231
0.25	331	74	0.224
1	552	130	0.236
2	308	65	0.211

a) For each of the four dose groups, calculate the logit of the estimated proportion. Plot the logit versus the dose of vitamin C.



b) Fit the logistic regression model using dose as the explanatory variable. Report the estimated model, the p-value from the deviance goodness-of-fit test, and the p-value for testing whether dose is required in the model.

```
y=cbind(c(267,74,130,65),c(1158-267,331-74,552-130,308-65))
vitC.logit=glm(y~dose,family=binomial(link=logit))

summary(vitC.logit)
Call: glm(formula = y ~ dose, family = binomial(link = logit))
Deviance Residuals:
           1           2           3           4
 -0.06856547 -0.2740503 0.5702126 -0.3530251

Coefficients:
                  Value Std. Error     t value
(Intercept) -1.20031411 0.06166784 -19.4641839
       dose -0.03464716 0.07112564  -0.4871261

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 0.7680264 on 3 degrees of freedom

Residual Deviance: 0.5295739 on 2 degrees of freedom
```

The fitted logistic model is:

$$\text{logit}(\hat{\pi}) = -1.20 - 0.035\,\text{dose}$$

The deviance goodness-of-fit statistic is 0.529 with a p-value of 0.77. This suggests that the model is appropriate (or the that test lacks power)

```
> 1 - pchisq(0.529, 2)
```

```
[1] 0.7675896
```

For testing whether dose is required in the model we have test statistic of -0.48 with a p-value of 0.63. This suggests that dose is not an important explanatory variable.

c) What can we conclude about the appropriateness of the logistic regression model? What evidence is there that the odds of a cold are associated with the dose of vitamin C?

From part (b) we have no evidence that the model is inappropriate. There is also no evidence that the odds of a cold are related to dose of vitamin C.

d) Why were both the doctors and patients unaware of the dose of Vitamin C contained in the tablets?

This was done to avoid the results being biased. If patients knew they were getting vitamin C they might respond to the idea of treatment rather than treatment itself (placebo effect). If doctors knew what doses each patient was getting it may affect their diagnosis – in a borderline case doctors might be less likely to diagnose cold for a patient getting a high dose of vitamin C.

Question 3 (ex from Chapter 21 of the class text)

Researchers in Kenya identified a cohort of more than 1,000 prostitutes who were known to be a major reservoir of sexually transmitted disease in 1985. It was determined that more than 85% of them were infected with HIV in February, 1986. The researchers then identified men who acquired a sexually transmitted disease from this group of women after the men sought treatment from a free clinic. The display below shows the subset of those men who did not test positive for HIV on their first visit and who agreed to participate in the study. The men are categorised according to whether they later tested positive for HIV during the study period, whether they had one or multiple sexual contacts with the prostitutes, and whether they were circumcised. Describe how the odds of testing positive are associated with number of contacts and with whether the male was circumcised.

The Lancet (1999): 403–07).

Display 21.22 Number of Kenyan men who tested positive for HIV, categorized according to two possible risk factors

	Single contact with prostitutes		Multiple contact with prostitutes	
	Circumcised	Uncircumcised	Circumcised	Uncircumcised
Tested positive for HIV	1	5	5	13
Number of men	46	27	168	52

19 Meta-Analysis of Breast Cancer and Lactation Studies

```
multpart=c(0,0,1,1)
circumcised=c(1,0,1,0)
Y=cbind(c(1,5,5,13),c(46-1,27-5,168-5,52-13))
HIV.logit=glm(Y~circumcised+multpart,family=binomial(link=logit))
summary(HIV.logit)
```

```
Call: glm(formula = Y ~ circumcised + multpart, family = binomial(link =
    logit))
Deviance Residuals:
          1           2           3           4
 0.03893972 -0.01880555 -0.01722502 0.01218789

Coefficients:
                Value Std. Error    t value
(Intercept) -1.4722967  0.4550932 -3.2351546
circumcised -2.3739845  0.4988156 -4.7592424
  multpart  0.3697799  0.5217637  0.7087115

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 27.09171 on 3 degrees of freedom

Residual Deviance: 0.0023152 on 1 degrees of freedom
```

0.3697799 ± 1.96×0.5217637

The fitted logistic model is:

$$\text{logit}(\hat{\pi}) = -1.47 - 2.37\text{circum} + 0.37\text{multpart}$$

circum=1 if circumcised, 0 otherwise

multpart=1 if had multiple partners, 0 otherwise

exp(0) = 1

It is estimated that the odds of being HIV positive were exp(0.37)=1.45 times greater for men with multiple partners (95% CI is 0.5 to 4). Since this CI contains 1, the data suggests that multiple partners is not significant. There is strong evidence that circumcised men were less likely to be HIV positive. It is estimated that the odds of being HIV positive for the circumcised men were 0.09 times the odds for uncircumcised men (95% CI 0.04 to 0.25).

CI of $\beta_{multpart}$