Statistical Modelling for Data Science Assignment 2

Chi Yin Wong

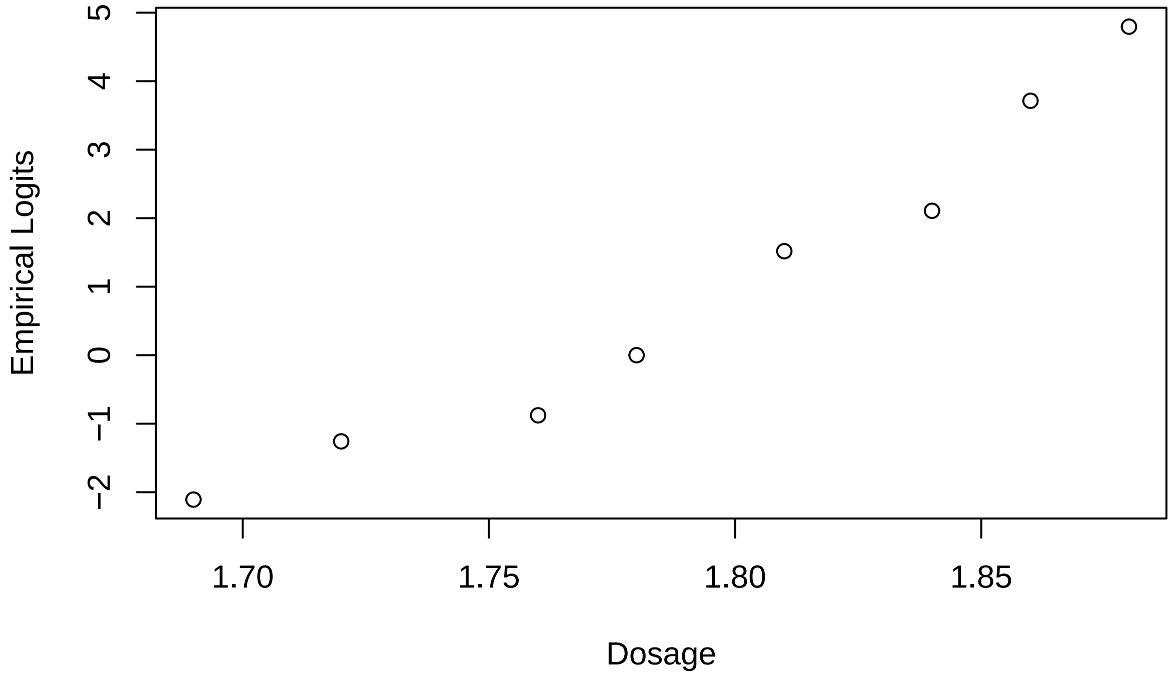
03/05/2021

Question 1

a

```
x = c(1.69,1.72,1.76,1.78,1.81,1.84,1.86,1.88)
n = c(59,60,62,56,63,59,62,60)
y = c(6,13,18,28,52,53,61,60)

plot(x, log((y+0.5)/(n-y+0.5)), xlab = "Dosage", ylab = "Empirical Logits")
```



The plot looks linear.

b

```
fit.1 = glm(y/n ~ x, family = binomial, weight = n)
summary(fit.1)
```

```
##
## Call:
## glm(formula = y/n ~ x, family = binomial, weights = n)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8986  -0.5475   0.9842   1.3315   1.7179
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.103      5.164  -11.64   <2e-16 ***
## x             33.934      2.903   11.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.202  on 7  degrees of freedom
## Residual deviance:  13.633  on 6  degrees of freedom
## AIC: 43.831
##
## Number of Fisher Scoring iterations: 4
```

The estimate of the intercept is -60.103. The estimate of the slope is 33.934.

c

From the summary above, the standard error of the slope is 2.903. Thus, a 95% confidence interval for the slope of the model is given by $(33.934 - 1.96 * 2.903, 33.934 + 1.96 * 2.903)$ which is equal to $(28.244, 39.624)$ rounded to 3 decimal places.

d

50% of the beetles dying is when $\pi = 0.5$. The estimate of the dosage that will kill 50% of the beetles is therefore given by

$$-\hat{\beta}_0/\hat{\beta}_1 = -(-60.103)/33.934$$

The estimate is 1.771 rounded to 3 decimal places.

e

When the dosage is incresed by 0.1, the odds of a beetle being killed increases by a factor of $e^{0.1*33.934} = 29.767$ rounded to 3 decimal places.

The approximate 95% confidence interval for the log-odds ratio is $(0.1 * 33.934 - 1.96 * 0.1 * 2.903, 0.1 * 33.934 + 1.96 * 0.1 * 2.903 = (2.824, 3.962)$.

Thus, the approximate 95% confidence interval for the factor by which the odds of being killed increase for a 0.1 increase in dosage is $(e^{2.824}, e^{3.962}) = (16.844, 52.583)$.

f

When dosage $= 1.8$, the estimated log odds of mortality is given by $-60.103 + 33.934 * 1.8 = 0.9782$. The standard error is found using the variance-covariance matrix og our linear model below.

```r
summary(fit.1)$cov.scaled
```

```
##             (Intercept)          x
## (Intercept)    26.66859 -14.986119
## x             -14.98612   8.426637
```

The standard error is $\sqrt{26.66859 + 1.8^2 * 8.426637 + 2 * (-14.986119)} = 4.899$.

The estimated probability is $e^{0.9782}/(1 + e^{0.9782}) = 0.72675$.

An approximate 95% confidence interval for the log-odds is $(0.9782 - 1.96 * 4.899, 0.9782 + 1.96 * 4.899) = (-8.624, 10.580)$.

Thus the approximate 95% confidence interval for the probability that a dosage of 1.8 is fatal is $(e^{-8.624}/(1 + e^{-8.624}), e^{10.580}/(1 + e^{10.580})) = (0.00018, 0.99997)$ to 5 decimal places.

g

The model fit.1 has 2 parameters and there are 8 groups. Thus the degrees of freedom are equal to 6. The critical value is then equal to $\chi^2_{0.95}(6) = 12.59159$.

```
# Chi square distribution with 6 df (95% quantile)
c.val.95 = qchisq(0.95,6)
c.val.95
```

```
## [1] 12.59159
```

```
# Residual deviance
deviance(fit.1)
```

```
## [1] 13.63338
```

The residual deviance statistic is equal to 13.63338 which is greater than the critical value. Thus, we reject the null hypothesis at the 5% level and conclude that there is no significant evidence that the model provides an adequete fit to the data.

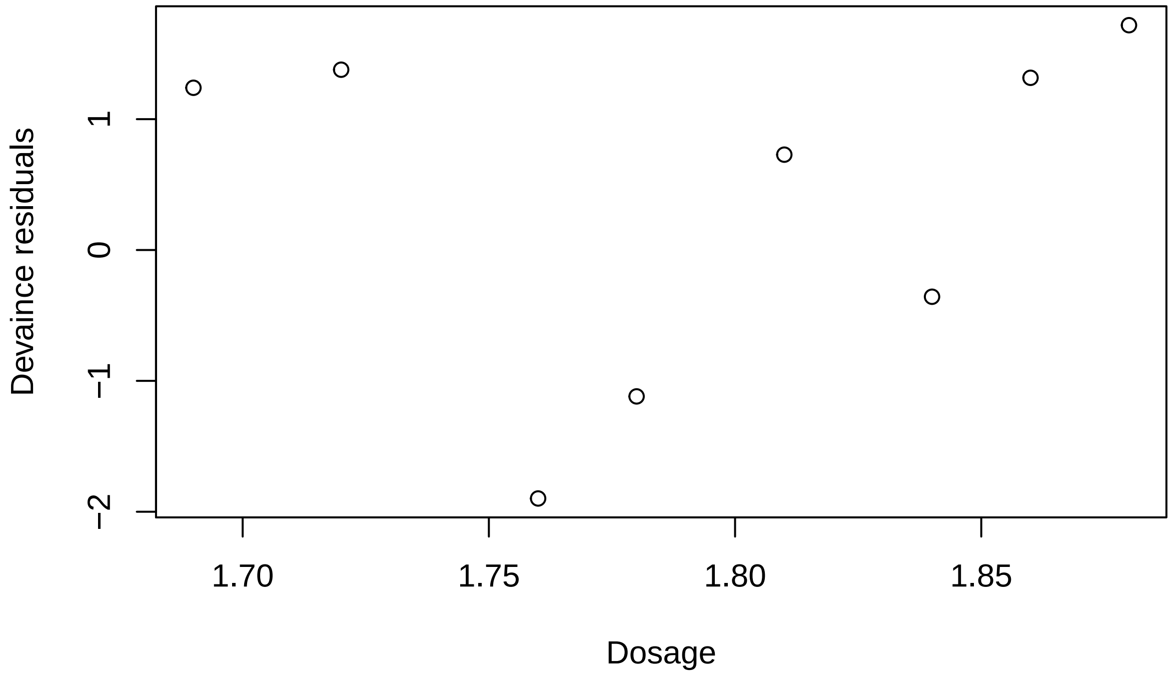
```
# Pearson deviance
sum(resid(fit.1, type = "pearson")^2)
```

```
## [1] 12.11328
```

The pearson chi square test statistic is equal to 12.11328 which is less than the critical value. Thus, we do not reject the null hypothesis at the 5% level and conclude that there is significant evidence that the model provides an adequete fit to the data.

h

```
fit.1.d.res = residuals(fit.1, type = "deviance")
plot(x, fit.1.d.res, xlab = "Dosage", ylab = "Devaince residuals")
```



Examining the plot, there appears to be no association between the deviance residauls and dosage. There is only one deviance residual that is particularly large in magnitude (group 3, dosage = 1.76).

i

```
fit.2 = glm(y/n ~ x + I(x^2), family = binomial, weight = n)
anova(fit.2, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y/n
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                       7    284.202
## x       1  270.569         6     13.633   <2e-16 ***
## I(x^2)  1    8.526         5      5.107   0.0035 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the analysis of deviance table, including a quadratic term (and consequently using 1 degree of freedom) will reduce the residual deviance by 8.526. This has a p-value of 0.0035 which is less than 0.05. This suggests that at a significance level of 5%, a quadratic logistic regression model will provide a better fit to the data than the straight line model.

Question 2

First we create a data frame that contains all the required data.

```r
sex = c(rep("male",12),rep("female",12))
education = c(rep(6:17), rep(6:17))
agree = c(25,27,75,29,32,36,115,31,28,9,15,3,17,26,91,30,55,50,190,17,18,7,13,3)
disagree = c(9,15,49,29,45,59,245,70,79,23,110,29,5,16,36,35,67,62,403,92,81,34,115,28)
total = c(34,42,124,58,77,95,360,101,107,32,125,32,22,42,127,65,122,112,593,109,99,41,128,31)
attitude.dat = data.frame(sex,education,agree,disagree,total)
```

a

We convert sex and education into factors and append them to our dataset. Then we fit a additive model with nominal main effects.

```r
sex.f = factor(sex)
attitude.dat$sex.f = sex.f
education.f = factor(education)
attitude.dat$education.f = education.f

fit.1 = glm(agree/total ~ sex.f + education.f, family = binomial, weight = total,
            data = attitude.dat)
summary(fit.1)
```

```
##
## Call:
## glm(formula = agree/total ~ sex.f + education.f, family = binomial,
##     data = attitude.dat, weights = total)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.82447  -0.32590   0.00052   0.30603   1.74278
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.08034    0.31309   3.451 0.000559 ***
## sex.fmale      0.03019    0.08740   0.345 0.729800
## education.f7  -0.55910    0.38270  -1.461 0.144030
## education.f8  -0.42588    0.33634  -1.266 0.205442
## education.f9  -1.17592    0.35770  -3.287 0.001011 **
## education.f10 -1.34462    0.34062  -3.948 7.89e-05 ***
## education.f11 -1.43565    0.33955  -4.228 2.36e-05 ***
## education.f12 -1.84536    0.31693  -5.823 5.79e-09 ***
## education.f13 -2.31131    0.34979  -6.608 3.91e-11 ***
## education.f14 -2.34261    0.35111  -6.672 2.52e-11 ***
## education.f15 -2.36410    0.41891  -5.643 1.67e-08 ***
## education.f16 -3.17924    0.36808  -8.637  < 2e-16 ***
## education.f17 -3.34706    0.52869  -6.331 2.44e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 337.90  on 23  degrees of freedom
## Residual deviance:  15.16  on 11  degrees of freedom
## AIC: 149.61
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
anova(fit.1, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: agree/total
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                           23     337.90
## sex.f        1     0.37        22     337.53   0.5431
## education.f 11   322.37        11      15.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, sex as a factor seems to be insignificant. We change the order of the variables around to see if sex is significant in the model.

```
fit.2 = glm(agree/total ~ education.f + sex.f, family = binomial, weight = total,
            data = attitude.dat)
anova(fit.2, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: agree/total
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                           23     337.90
## education.f 11   322.62        12      15.28   <2e-16 ***
## sex.f        1     0.12        11      15.16   0.7298
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sex as a factor is insignificant according to the analysis of deviance table in both models, meaning that if education as a factor is a predictor in the model, then we can exclude sex as a predictor. Now we test whether the effects of sex and years are strictly additive by testing to see if the model with an interaction term improves the fit significantly.

```
fit.3 = glm(agree/total ~ education.f*sex.f, family = binomial, weight = total,
            data = attitude.dat)
anova(fit.3, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
```

```
## Response: agree/total
##
## Terms added sequentially (first to last)
##
##
##                   Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                                23     337.90
## education.f       11   322.62        12      15.28   <2e-16 ***
## sex.f              1     0.12        11      15.16   0.7298
## education.f:sex.f 11    15.16         0       0.00   0.1753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the analysis of deviance table, the addition of an interaction term will reduce the residual deviance to 0 but also use all the remaining degrees of freedom, causing the model to become saturated. The p-value is 0.1753, suggesting that at the 5% level, adding an interaction term does not improve the fit sigificantly. Thus, we can conclude that the effects of gender and years of education as factors are strictly additive.

b

First we fit the model with sex as a factor and years of education as a numerical variable.

```
fit.4 = glm(agree/total ~ sex.f + education, family = binomial, weight = total,
            data = attitude.dat)
anova(fit.4, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: agree/total
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         23     337.90
## sex.f      1     0.37        22     337.53   0.5431
## education  1   312.45        21      25.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the analysis of deviance table, sex as a factor has a p-value of 0.5431, so it is not significant, but education is highly significant with a very small p-value. We rearrange the order they are sequentially added into the model to see if there is a difference.

```
fit.5 = glm(agree/total ~ education + sex.f, family = binomial, weight = total,
            data = attitude.dat)
anova(fit.5, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: agree/total
```

```
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                          23     337.90
## education  1  312.666        22      25.24   <2e-16 ***
## sex.f      1    0.149        21      25.09   0.6995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We get the same result (sex.f has a p-value of 0.6995), so we can remove sex.f from the model as it does not significantly improve our fit when we already have education as a predictor. Now we see if adding education as a factor after education will further improve the fit of our model.

```
fit.6 = glm(agree/total ~ education + factor(education), family = binomial, weight = total,
            data = attitude.dat)
anova(fit.6, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: agree/total
##
## Terms added sequentially (first to last)
##
##
##                   Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                                  23     337.90
## education          1  312.666        22      25.24   <2e-16 ***
## factor(education) 10    9.957        12      15.28   0.4443
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the analysis of deviance table, education as a factor has a p-value of 0.4443, so we can conclude that it does not make a significant improvement to the model. Next we try fit a model with education as a quadratic term.

```
fit.7 = glm(agree/total ~ education + I(education^2), family = binomial, weight = total,
            data = attitude.dat)
anova(fit.7, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: agree/total
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                               23     337.90
## education       1  312.666        22      25.24   <2e-16 ***
## I(education^2)  1    0.014        21      25.22   0.9042
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the analysis of deviance table, education as a quadratic term has a p-value of 0.9042. It does not significantly improve the fit of the model. Thus, the best logistic regression model is the model that only contains education as a numerical variable.

```r
fit.8 = glm(agree/total ~ education, family = binomial, weight = total, data = attitude.dat)
anova(fit.8, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: agree/total
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         23     337.90
## education  1   312.67        22      25.24 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(fit.8)
```

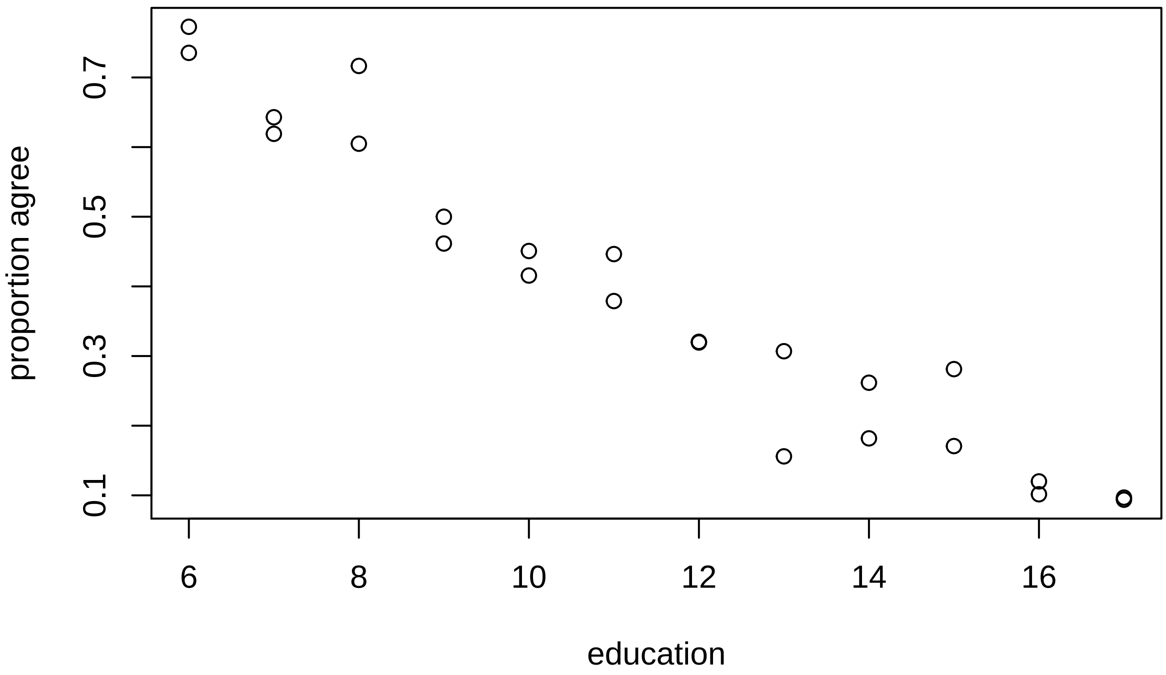
```
##
## Call:
## glm(formula = agree/total ~ education, family = binomial, data = attitude.dat,
##     weights = total)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5708  -0.5734  -0.1063   0.2246   2.4513
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.88286    0.21454   13.44   <2e-16 ***
## education   -0.30297    0.01855  -16.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 337.903  on 23  degrees of freedom
## Residual deviance:  25.236  on 22  degrees of freedom
## AIC: 137.68
##
## Number of Fisher Scoring iterations: 4
```

The analysis of deviance table reinforces this, as the p-value is extremely small, suggesting that education as a numerical variable provides a very good fit for the data. The odds ratio is $e^{-0.30297} = 0.739$ (to 3 decimal places). This mean that the odds of agreeing with the statement "Women should take care of running their homes and leave the running of the country up to men" decreases by a factor of 0.739 for each increase in education by one year.

c

After a through analysis in part a and b, it seems that the model that only uses the numerical variable education provides the best fit for the data. From the analysis of deviance table (fit.8), we see that with 1 degree of freedom, education reduces the residual deviance by 312.67. We plot years of education against the proportion of people who agree to get a visual relationship and determine whether it is consistent with our results.

```
plot(education, agree/total, xlab = "education", ylab = "proportion agree")
```



There is clearly a negative linear association between education and the proportion of peopole who agree with the statement, which is consistent with our interpretation in part b. Furthermore, we compare the analysis of deviance table of fit.8 to fit.4 (where sex as a factor was included in the model) and we see that sex.f uses 1 degree of freedom but only reduces the residual deviance by 0.37, so it does not significantly improve the fit of the model. Overall, it is clear that the model most appropriate for our dataset is the linear model with only education as a numerical variable.

Question 3

a

Firstly we create a data frame that contains all the data to be used to fit log-linear models.

```
d.race = c(rep(c("white", "white", "black", "black"),2))
v.race = c(rep(c("white", "black"),4))
sentence = c(rep("yes",4), rep("no",4))
count = c(19,0,11,6,132,9,52,97)
death.dat = data.frame(d.race,v.race,sentence,count)
```

The hypothesis test for each scenario is conducted in the code below. Test 1 (fit.1) is associated with part i, test 2(fit.2) with part ii and so on.

```
# Test 1: all three factors are mutually independent
fit.1 = glm(count ~ d.race + v.race + sentence, family = poisson, data = death.dat)
summary(fit.1)
```

```
##
## Call:
## glm(formula = count ~ d.race + v.race + sentence, family = poisson,
##     data = death.dat)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
##  1.9881  -3.4843  -0.3023  -0.1196   3.7542  -7.0237  -5.0100   5.7623
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.92657    0.11075  35.455  < 2e-16 ***
## d.racewhite -0.03681    0.11079  -0.332     0.74
## v.racewhite  0.64748    0.11662   5.552 2.83e-08 ***
## sentenceyes -2.08636    0.17671 -11.807  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 395.92  on 7  degrees of freedom
## Residual deviance: 137.93  on 4  degrees of freedom
## AIC: 181.61
##
## Number of Fisher Scoring iterations: 5
```

```r
deviance(fit.1)
```

```
## [1] 137.9294
```

```r
# Test 2: sentence is independent of both the defendant and the victims race
fit.2 = glm(count ~ sentence + d.race*v.race, family = poisson, data = death.dat)
summary(fit.2)
```

```
##
## Call:
## glm(formula = count ~ sentence + d.race * v.race, family = poisson,
##     data = death.dat)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
##  0.5569  -1.4099   1.4118  -1.7531  -0.2012   0.3443  -0.5467   0.5561
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)               4.5177     0.1004  44.976  < 2e-16 ***
## sentenceyes              -2.0864     0.1767 -11.807  < 2e-16 ***
## d.racewhite              -2.4375     0.3476  -7.013 2.34e-12 ***
## v.racewhite              -0.4916     0.1599  -3.074  0.00212 **
## d.racewhite:v.racewhite   3.3116     0.3786   8.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 395.9153  on 7  degrees of freedom
## Residual deviance:   8.1316  on 3  degrees of freedom
## AIC: 53.813
##
## Number of Fisher Scoring iterations: 4
```

```
deviance(fit.2)
```

```
## [1] 8.131611
```

```
# Test 3: given d.race, sentence is independent of v.race
fit.3 = glm(count~ sentence*d.race + v.race*d.race, family = poisson, data = death.dat)
summary(fit.3)
```

```
##
## Call:
## glm(formula = count ~ sentence * d.race + v.race * d.race, family = poisson,
##     data = death.dat)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
##  0.24994  -1.46202   1.62523  -1.52514  -0.09277   0.37142  -0.61322   0.46922
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)               4.5267     0.1020  44.396  < 2e-16 ***
## sentenceyes              -2.1707     0.2560  -8.480  < 2e-16 ***
## d.racewhite              -2.4559     0.3498  -7.021  2.2e-12 ***
## v.racewhite              -0.4916     0.1599  -3.074  0.00212 **
## sentenceyes:d.racewhite   0.1664     0.3539   0.470  0.63821
## d.racewhite:v.racewhite   3.3116     0.3786   8.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 395.9153  on 7  degrees of freedom
## Residual deviance:   7.9102  on 2  degrees of freedom
## AIC: 55.592
##
## Number of Fisher Scoring iterations: 4
```

```
deviance(fit.3)
```

```
## [1] 7.910161
```

```
# Test 4: given v.race, sentence is independent of d.race
fit.4 = glm(count~ sentence*v.race + d.race*v.race, family = poisson, data = death.dat)
summary(fit.4)
```

```
##
## Call:
## glm(formula = count ~ sentence * v.race + d.race * v.race, family = poisson,
##     data = death.dat)
##
## Deviance Residuals:
```

```
##        1         2         3         4         5         6         7         8
## -0.47967  -0.98198   0.70243   0.20237   0.18976   0.16368  -0.29660  -0.04887
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)               4.5797     0.1011  45.314  < 2e-16 ***
## sentenceyes              -2.8717     0.4196  -6.843 7.75e-12 ***
## v.racewhite              -0.5876     0.1639  -3.586 0.000336 ***
## d.racewhite              -2.4375     0.3476  -7.013 2.34e-12 ***
## sentenceyes:v.racewhite   1.0579     0.4635   2.282 0.022471 *
## v.racewhite:d.racewhite   3.3116     0.3786   8.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 395.9153  on 7  degrees of freedom
## Residual deviance:   1.8819  on 2  degrees of freedom
## AIC: 49.563
##
## Number of Fisher Scoring iterations: 4
```

```
deviance(fit.4)
```

```
## [1] 1.881895
```

Based on the chi-square goodness of fit tests, the only 2 models that provide an adequte fit to the data are fit.3 [SD][VD] and fit.4[SV][DV]. Since both have degrees of freedom of 2, but fit.4 has a lower residual deviance, we conclude that fit.4 [SV][DV] is the most appropriate model, and that the sentence depends on the victim's race, but given the victim's race, is independent of the defendant's race.

b

Firstly we create a data frame so that we can use for logistic regression.

```
d.race.logit = factor(c("white","white","black","black"))
v.race.logit = factor(c("white", "black"))
yes = c(19,0,11,6)
no = c(132,9,52,97)
total = yes+no
death.dat.logit = data.frame(d.race.logit,v.race.logit,yes,total)
```

Then, for each scenario, we fit a logistic regression model and use the anova function to test the hypotheses. In the first scenario, all three factors are mutually independent is the same as fitting a logistic regression model with an interaction term, and then using an anova test to determine whether the interaction term improves the fit of the model significantly.

```
fit.5 = glm(yes/total~d.race.logit*v.race.logit, family = binomial, weight = total,
            data = death.dat.logit)
anova(fit.5, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: yes/total
```

```
##
## Terms added sequentially (first to last)
##
##
##                            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                                           3     8.1316
## d.race.logit                1   0.2215         2     7.9102 0.637937
## v.race.logit                1   7.2094         1     0.7007 0.007252 **
## d.race.logit:v.race.logit   1   0.7007         0     0.0000 0.402535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the analysis of deviance table, we see that the interaction term has a p-value of 0.403, so we conclude that the interaction term is insignificant, and this model does not provide an adequete fit. In the second scenario, sentence is independent of both the defendant and the victim's race is the same as fitting a logistic regression model with only the intercept term.

```
fit.6 = glm(yes/total~1, family = binomial, weight = total, data = death.dat.logit)
anova(fit.6, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: yes/total
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                     3     8.1316
```

From the analysis of deviance table, we see that the residual devaince is 8.1316 which is quite large. Thus, we conclude that the model does not provide an adequte fit. In the third scenario, given the defendants race, sentence is independent of the victim's race is the same as a logistic regression model with only the defendent variable.

```
fit.7 = glm(yes/total~d.race.logit, family = binomial, weight = total,
            data = death.dat.logit)
anova(fit.7, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: yes/total
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                             3     8.1316
## d.race.logit  1  0.22145         2     7.9102   0.6379
```

From the analysis of deviance table, we see that the p-value is 0.6379, so we conclude that the model does not provide an adequete fit for the data. In the fourth scenario, given the victim's race sentence is independent of the defendent's race is the same as a logistic regression model with only the victim variable.

```
fit.8 = glm(yes/total~v.race.logit, family = binomial, weight = total,
            data = death.dat.logit)
anova(fit.8, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: yes/total
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                             3     8.1316
## v.race.logit  1   6.2497         2     1.8819  0.01242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the analysis of deviance table, we see that the p-value is 0.01242, which is significant at the 5% level. Thus, we conclude that there is evidence to support that this model does provide an adequete fit for the data. From all the test conducted using the logistic regression model, we can conclude that both the log-linear model approach and the logistic regression approach arrive at the same conclusion: the model where given the victim's race, sentence is independent of the defendant's race provides the best fit for the data.