Practice 7

1. The data below come from a survey of opinion on the Vietnam war conducted among 1st – 3rd year undergraduate students at the University of North Carolina in 1967.

   The policies listed were:

   A – defeat power of Vietnam by widespread bombing and land invasion

   B – follow the present policy

   C – withdraw troops to strong points and open negotiations on elections involving the Vietcong

   D – immediate withdrawal of all U.S. troops.

		Opinion (op)			
Gender	Year	A	B	C	D
Male	1	175	116	131	17
	2	160	126	135	21
	3	132	120	154	29
Female	1	13	19	40	5
	2	5	9	33	3
	3	22	29	110	6

	Model	scaled deviance	residual df
1	gender+year+op	181.42	17
2	gender*year+op	120.05	15
3	gender*op+year	78.88	14
4	gender+year*op	157.33	11
5	gender*year+gender*op	17.51	12
6	gender*year+year*op	95.96	9
7	gender*op+year*op	54.79	8
8	gender*year+gender*op+year*op	5.70	6

   (a) Which of the three factors, `Opinion`, `Gender` and `Year`, can serve as the response variable? Why? Based on your answer, which of the eight models is (or are) not worth to consider? Why not?

   (b) Which of the following conclusions are reasonable for the data? Why or why not?

      i. Opinion is independent of year level and gender.

      ii. Given the year level, opinion is independent of gender.

      iii. Given gender, opinion is independent of year level.

      iv. The ratio of the odds between any two opinions (B and C say) for males and females is the same for each year level.

   Which, if any, of these conclusions do you consider to be the most appropriate based on the model deviance output provided? Justify your answer.

(c) It has been suggested that the most appropriate model found aboveit might be improved by including year as a variable (as *yr*) rather than as a factor in some way. Write down the form of such a model which is likely to provide an adequate fit to the data, and give an interpretation of the model.

(d) Write down the deviance and the degrees of freedom that would be obtained by fitting the model `year+op` to the 2–way table obtained by collapsing over `gender`. State, with reasons, whether collapsing over `gender` would be a reasonable thing to do.

2. The following data were obtained from a study of coronary heart disease, where `N` is the total number of subjects in each group and `Y` is the number diagnosed with coronary heart disease. The factor `CHOL` refers to serum cholesterol in mg/100cc where:

$$1 = < 200,\ 2 = 200 - 219,\ 3 = 220 - 259,\ 4 = 260+$$

while the factor `BP` refers to blood pressure in mm of mercury where:

$$1 = < 127,\ 2 = 127 - 146,\ 3 = 147 - 166,\ 4 = 167+$$

		BP			
CHOL		1	2	3	4
1	Y	2	3	3	4
	N	119	124	50	26
2	Y	3	2	0	3
	N	88	100	43	23
3	Y	8	11	6	6
	N	127	220	74	49
4	Y	7	12	11	11
	N	74	111	57	44

Four models have been fitted to these data, R output for which is given below.

```
> Y <- c(2, 3, 3, 4, 3, 2, 0, 3, 8, 11, 6, 6, 7, 12, 11, 11)
> N <- c(119, 124, 50, 26, 88, 100, 43, 23, 127, 220, 74, 49, 74,
+     111, 57, 44)
> BP <- factor(rep(1:4, 4))
> CHOL <- factor(rep(1:4, rep(4, 4)))
> fit.1 <- glm(Y/N ~ 1, weights = N, family = "binomial")
> summary(fit.1)

Call:
glm(formula = Y/N ~ 1, family = "binomial", weights = N)
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.67546  -1.63956   0.06465   1.37102   3.74137

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.5987     0.1081  -24.05   <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: 58.726  on 15  degrees of freedom
Residual deviance: 58.726  on 15  degrees of freedom
AIC: 111.83

Number of Fisher Scoring iterations: 5

> fit.2 <- glm(Y/N ~ CHOL, weights = N, family = "binomial")
> summary(fit.2)

Call:
glm(formula = Y/N ~ CHOL, family = "binomial", weights = N)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-1.6589861  -1.0203129   0.0009951   1.1270950   2.3674007

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.2419     0.2943 -11.017  < 2e-16 ***
CHOL2        -0.1839     0.4644  -0.396   0.6920
CHOL3         0.5914     0.3480   1.699   0.0893 .
CHOL4         1.4543     0.3392   4.287 1.81e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58.726  on 15  degrees of freedom
Residual deviance: 26.805  on 12  degrees of freedom
AIC: 85.909

Number of Fisher Scoring iterations: 5

> fit.3 <- glm(Y/N ~ BP, weights = N, family = "binomial")
> summary(fit.3)
```

```
Call:
glm(formula = Y/N ~ BP, family = "binomial", weights = N)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8361  -1.0499  -0.3808   0.8645   2.4265

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.96527    0.22930 -12.932  < 2e-16 ***
BP2          0.03028    0.30032   0.101   0.9197
BP3          0.64289    0.32784   1.961   0.0499 *
BP4          1.37264    0.32050   4.283 1.85e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58.726  on 15  degrees of freedom
Residual deviance: 35.163  on 12  degrees of freedom
AIC: 94.267

Number of Fisher Scoring iterations: 5

> fit.4 <- glm(Y/N ~ CHOL + BP, weights = N, family = "binomial")
> summary(fit.4)

Call:
glm(formula = Y/N ~ CHOL + BP, family = "binomial", weights = N)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.89259  -0.34946  -0.02072   0.52307   0.99198

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.48194    0.34865  -9.987  < 2e-16 ***
CHOL2       -0.20798    0.46641  -0.446 0.655663
CHOL3        0.56223    0.35080   1.603 0.108998
CHOL4        1.34412    0.34297   3.919 8.89e-05 ***
BP2         -0.04146    0.30365  -0.137 0.891393
BP3          0.53236    0.33240   1.602 0.109251
BP4          1.20042    0.32689   3.672 0.000240 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

```
    Null deviance: 58.7262  on 15  degrees of freedom
Residual deviance:  8.0762  on  9  degrees of freedom
AIC: 73.18

Number of Fisher Scoring iterations: 4
```

The data have been analysed using logistic regression models as shown above. An alternative would have been to use log-linear models in a 3-way contingency table with factors `CHOL1`, `BP1` and `CHD`, where `CHD` is a factor with 2 levels indicating whether or not subjects have coronary heart disease. For each of the four (logistic regression) models given in the `R` output, specify the equivalent log-linear model (eg `CHOL1 + BP1 + CHD`). Also analyze these log-linear models using `R`, and compare the results with that of the logistic models.