Practice 5

1. If the distribution of Y is a member of the exponential family, and is in 'canonical' form then

$$\ln f(y|\theta, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

where θ is the natural parameter and ϕ is the dispersion parameter.

We have the following properties

$$\mathbb{E}(Y) = b'(\theta) \quad \text{and} \quad \text{Var}(Y) = a(\phi)b''(\theta)$$

Show that these results hold for each of the following distributions:

(a) $Y \overset{d}{=} \text{Bin}(n,\, p)$;

(b) $Y \overset{d}{=} \text{Poi}(\lambda)$.

2. The following data are on the model

$$Y_i \overset{d}{=} \text{Poi}(\lambda_i) \quad \text{where } \ln \lambda_i = \alpha + \beta x_i$$

x	32.7	38.3	39.8	30.0	34.3	36.3	32.5	40.0	30.4	28.2
y	5	10	12	3	6	8	4	12	3	3

(a) Find the MLEs of α and β and give their standard errors.

(b) Plot the data and the fitted model on a suitable graph.

3. 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
```

(a) Which of the four models is “best”? Give details of any formal tests that you use in reaching your decision.

(b) Describe briefly (no calculations required) what your chosen model says, if anything, about the relationships between:

  i. coronary heart disease and serum cholesterol levels;

  ii. coronary heart disease and blood pressure;

  iii. serum cholesterol levels and blood pressure.

(c) The model with `CHOL` and `BP` included as variables, rather than as factors, was fitted to the data and resuted in a scaled deviance of 14.847. What conclusions do you draw from this? [Give details of any formal tests that you use.]