Practice 5 Solutions

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

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

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

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

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

- (a). The *pdf* (or *pmf*) for binomial distribution is $f_Y(y|p) = \binom{n}{y}p^y(1-p)^{n-y}$. Therefore

$$\ln f_Y(y|p) = \ln\binom{n}{y} + y\ln p + (n-y)\ln(1-p) = y\ln\frac{p}{1-p} + n\ln(1-p) + \ln\binom{n}{y}.$$

According to the specification above, $\theta = \ln\frac{p}{1-p}$ and accordingly $p = \frac{e^\theta}{1+e^\theta}$. So the function $b(\theta)$ must be $b(\theta) = -n\ln(1-p) = n\ln(1+e^\theta)$. Then $b'(\theta) = \frac{ne^\theta}{1+e^\theta} = np = \mathbb{E}(Y)$, and $b''(\eta) = \frac{ne^\theta}{(1+e^\theta)^2} = np(1-p) = \mathrm{Var}(Y)$. (Note $a(\phi) = 1$ here.)

- (b). The *pdf* (or *pmf*) for Poisson distribution is $f_Y(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$. Therefore

$$\ln f_Y(y|\lambda) = -\lambda + y\ln\lambda - \ln y!$$

According to the specification above, $\theta = \ln\lambda$, so the function $b(\theta)$ must be $b(\theta) = e^\theta$. Then $b'(\theta) = b''(\theta) = e^\theta = \lambda = \mathbb{E}(Y) = \mathrm{Var}(Y)$. (Note $a(\phi) = 1$ here.)

2. The following data are on the model

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

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.

- The MLEs and their standard errors are $\hat\alpha = -2.8177$, $\hat\beta = 0.1333$, $se(\hat\alpha) = 1.202$ and $se(\hat\beta) = 0.0329$. All these results can be obtained by using the R commands `glm.2a=glm(y~x, family=poisson)` and `summary(glm.2a)`.

- Alternatively, we can derive the method of scoring formula by ourselves and write our own R function to implement it. The results obtained are the same.

  $$\begin{aligned}
  \ell(\alpha,\beta) &= -\sum_{i=1}^{n}[e^{\alpha+\beta x_i} - \alpha y_i - \beta x_i y_i + \ln y_i!] \quad \text{(log-likelihood)} \\
  \mathbf{s}(\alpha,\beta) &= \frac{\partial \ell}{\partial(\alpha,\beta)^t} = \begin{pmatrix} -\sum_{i=1}^{n}[e^{\alpha+\beta x_i} - y_i] \\ -\sum_{i=1}^{n}[e^{\alpha+\beta x_i} x_i - x_i y_i] \end{pmatrix} \quad \text{(score function)} \\
  I(\alpha,\beta) &= E\left[\frac{-\partial^2 \ell}{\partial(\alpha,\beta)^t \partial(\alpha,\beta)}\right] = \begin{bmatrix} \sum_{i=1}^{n} e^{\alpha+\beta x_i} & \sum_{i=1}^{n} e^{\alpha+\beta x_i} x_i \\ \sum_{i=1}^{n} e^{\alpha+\beta x_i} x_i & \sum_{i=1}^{n} e^{\alpha+\beta x_i} x_i^2 \end{bmatrix} \quad \text{(Fisher informat} \\
  \begin{pmatrix} \alpha \\ \beta \end{pmatrix}_{k+1} &= \begin{pmatrix} \alpha \\ \beta \end{pmatrix}_{k} + I(\alpha_k,\beta_k)^{-1}\mathbf{s}(\alpha_k,\beta_k) \quad \text{(method of scoring iteration)}
  \end{aligned}$$

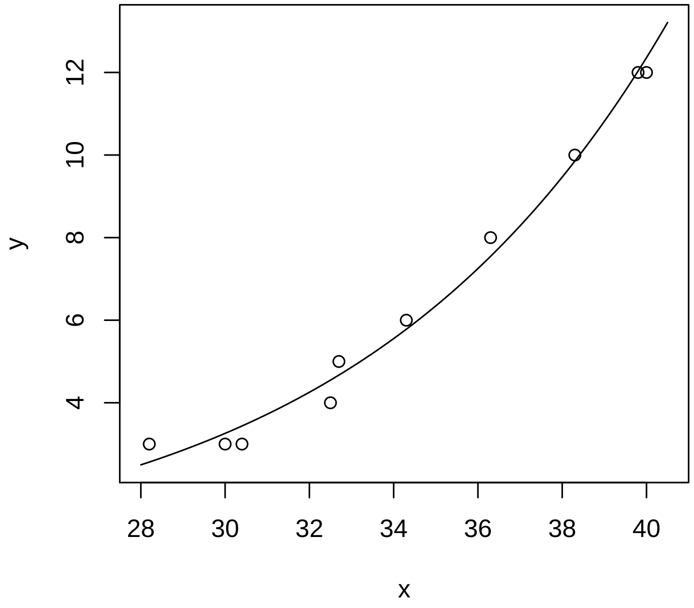
  The initial estimate of (α,β) is taken to be the least squares estimate obtained by fitting $\ln y = \alpha_0 + \beta_0 x$, which is (-2.8306, 0.1336).

-
  ```
  > x
   [1] 32.7 38.3 39.8 30.0 34.3 36.3 32.5 40.0 30.4 28.2
  > y
   [1]  5 10 12  3  6  8  4 12  3  3

  > glm.2a=glm(y~x, family=poisson)
  > summary(glm.2a)

  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept) -2.81772    1.20204  -2.344   0.0191 *
  x            0.13328    0.03294   4.047 5.19e-05 ***
  ---

  > lm(log(y)~x)$coef

  (Intercept)           x
   -2.8306010   0.1335509

  > mleT4q2.f

  function(y,x,theta0,iter=20){
          theta.mat=matrix(0,iter+1,2)
          theta.mat[1,]=theta0
          #print(theta.mat)
          for(k in 1:iter){
                  temp=exp(theta.mat[k,1]+theta.mat[k,2]*x)
                  #print(sum(temp))
                  u=c(-sum(temp-y),-sum((temp-y)*x))
                  #print(u)
          H.mat=matrix(c(sum(temp),sum(temp*x),sum(temp*x),sum(temp*(x^2))),2,2)
          #print(H.mat)
          theta.mat[k+1,]=theta.mat[k,]+solve(H.mat)%*%u}
          result=list(est=theta.mat, se=diag(solve(H.mat))^0.5)
          return(result)
  }

  > mleT4q2.f(y,x, c(-2.8306,0.1336),iter=4)

  $est
            [,1]      [,2]
  ```

```
[1,] -2.830600 0.1336000
[2,] -2.817695 0.1332804
[3,] -2.817716 0.1332809
[4,] -2.817716 0.1332809
[5,] -2.817716 0.1332809

$se
[1] 1.20203736 0.03293577
```

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

```
> curve(exp(-2.8177+0.1333*x), 28, 40.5, xlab="x", ylab="y")
> points(x,y)
```



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.

- The best model is `CHOL + BP` with

$$\text{logit}(\hat{p}) = -3.482 - 0.208\text{CHOL2} + 0.562\text{CHOL3} + \cdots + 1.200\text{BP4}.$$

  This is the only one of the four models which provides an adequate fit to the data. Specifically, the residual deviance of the model is 8.0762 with 9 degrees of freedom, and p-value =0.5265 based on the χ^2 test of adequacy.

- The model `CHOL+BP` means that the risk of `CHD` (coronary heart disease) depends on both `CHOL` and `BP`, and that the effects are additive on the logit scale.
- Also `BP` is significant after `CHOL` ($\Delta D = 26.805 - 8.0762 = 18.73$ on 3 df, with p-value of 0.0003); and `CHOL` is significant after `BP` ($\Delta D = 35.163 - 8.0762 = 27.09$ on 3 df, with p-value of 5.6×10^{-6}).

(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.

- The risk, odds and log-odds of `CHD` tend to increase with increasing `CHOL` and/or `BP`.

  i. `CHD` increases as `CHOL` increases.

  ii. `CHD` increases as `BP` increases.

  iii. The model provides no information as to any association between `CHOL` and `BP`.

(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.]

- Denote M_1 as the model `CHOL+BP`, and M_2 as the new model where `CHOL` and `BP` are treated as variables.
  The change in scaled deviance between M_1 and M_2 is $14.847 - 8.076 = 6.7708$ on 4 df, which is not significant (p-value$=$ 0.1485). Therefore the simpler model M_2 is not significantly worse than the more complicated one M_1. Also the model M_2 provides an adequate fit to the data: $D = 14.847$ on 13 df providing a p-value of 0.317.
- We can conclud that there is a simple linear trend between `CHD` and (`CHOL` and `BP`) on the logit scale.