Practice 6 Solutions

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
```

1. 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}).

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

   (a) coronary heart disease and serum cholesterol levels;

   (b) coronary heart disease and blood pressure;

   (c) serum cholesterol levels and blood pressure.

   - The risk, odds and log-odds of `CHD` tend to increase with increasing `CHOL` and/or `BP`.
     (a) `CHD` increases as `CHOL` increases.
     (b) `CHD` increases as `BP` increases.
     (c) The model provides no information as to any association between `CHOL` and `BP`.

3. The model with `CHOL` and `BP` included as variables, rather than as factors, was fitted to the data and resulted 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 conclude that there is a simple linear trend between `CHD` and (`CHOL` and `BP`) on the logit scale which does not provide significant evidence of inadequacy of fit.

4. Use `R` to fit the logistic model specified in question 3. Verify the conclusions drawn in the previous question. Also use the Pearson deviance to test the adequacy of this model.

- The following R output confirms the conclusions from 3.
- The Pearson deviance of the linear trend model in 3 equals 13.429 corresponding to $\chi^2(13)$ distribution. The resultant p-value is 0.415.

```
BP.v <- rep(1:4, 4);  CHOL.v <- rep(1:4, rep(4, 4))
fit.5 <- glm(Y/N ~ CHOL.v + BP.v, weights = N, family = "binomial")
summary(fit.5)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.0916     0.4428 -11.499  < 2e-16 ***
CHOL.v        0.5300     0.1166   4.547 5.45e-06 ***
BP.v          0.4405     0.1091   4.037 5.41e-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: 14.847  on 13  degrees of freedom
AIC: 71.951

summary(fit.5)$cov.scaled

            (Intercept)       CHOL.v         BP.v
(Intercept)  0.19604815 -0.038957319 -0.026256932
CHOL.v      -0.03895732  0.013585723 -0.001033497
BP.v        -0.02625693 -0.001033497  0.011902642

anova(fit.5, 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                      15     58.726
CHOL.v  1   27.832        14     30.894 1.323e-07 ***
BP.v    1   16.047        13     14.847 6.180e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. List both the deviance residuals and Pearson residuals of the model in 3 in a matrix form. Then comment on these residuals.

- Most residuals (both deviance and Pearson ones) are within -2 and $+2$ except a less than -2 deviance residual when `CHOL = 2` and `BP =3`, and a greater than 2 Pearson residual when `CHOL=1` and `BP=4`. This provides no significant evidence against the adequacy of the model fit.
- The two types of residuals are very close to each other, with the Pearson residuals tend to be bigger.

```
matrix(resid(fit.5, type="deviance"), 4, 4, byrow=T)

           [,1]        [,2]       [,3]       [,4]
[1,] 0.07259681 -0.02814454  0.7652350  1.7731758
[2,] 0.40525514 -1.17498731 -2.3532771  0.5750028
[3,] 0.94010870 -1.09879115 -0.6012941 -0.5419821
[4,] 0.66157521 -0.06327878  0.6398534  0.3169260

matrix(resid(fit.5, type="pearson"), 4, 4, byrow=T)

           [,1]        [,2]       [,3]       [,4]
[1,] 0.07321581 -0.02807163  0.8289876  2.1168841
[2,] 0.42194772 -1.06129659 -1.6911694  0.6055837
[3,] 0.99556263 -1.05055237 -0.5822558 -0.5277568
[4,] 0.68839456 -0.06312047  0.6559412  0.3200935

> diff=resid(fit.5, type="pearson")-resid(fit.5, type="deviance")
> matrix(diff, 4, 4, byrow=T)

            [,1]         [,2]       [,3]        [,4]
[1,] 0.000619007 7.290589e-05 0.06375260 0.343708316
[2,] 0.016692578 1.136907e-01 0.66210769 0.030580852
[3,] 0.055453936 4.823877e-02 0.01903821 0.014225336
[4,] 0.026819351 1.583157e-04 0.01608788 0.003167484
```

6. Fit a logistic regression model for coronary heart disease (CHD) which includes `CHOL` and `BP`, both as variables, plus their interaction term. Test the significance of this interaction term in the model by both the Wald test and the likelihood ratio test. Then compare this model with the linear trend model in question 3, and draw a conclusion.

   - The Wald test statistic for testing the significance of `CHOL.v:BP.v` interaction effect equals -0.949, with p-value of 0.34239. The likelihood ratio test statistic for this test equals the reduction of deviance due to `CHOL.v:BP.v`, which equals 0.900 with p-value 0.3428. Both tests suggests no significant evidence of the `CHOL.v:BP.v` interaction effect.
   - Thus, there is no significant difference between the two logistic models `fit.5` and `fit.6` in regard to goodness of fit. Therefore we prefer the simpler model which is the linear trend model `fit.5`.

```
fit.6 <- glm(Y/N ~ (CHOL.v + BP.v)^2, weights = N, family = "binomial")
summary(fit.6)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4263  -0.5886   0.3029   0.6775   1.2712

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9045     0.9900  -5.964 2.46e-09 ***
CHOL.v        0.7962     0.3082   2.583  0.00979 **
BP.v          0.7695     0.3632   2.119  0.03410 *
CHOL.v:BP.v  -0.1073     0.1130  -0.949  0.34239
---
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: 13.947  on 12  degrees of freedom
AIC: 73.051

anova(fit.6, 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                           15     58.726
CHOL.v       1   27.832        14     30.894 1.323e-07 ***
BP.v         1   16.047        13     14.847 6.180e-05 ***
CHOL.v:BP.v  1    0.900        12     13.947    0.3428
```

7. Using the model in 3, estimate the odds ratio of CHD when `CHOL` increases by one level and `BP` is kept unchanged. Also find an approximately 95% confidence interval for this odds ratio.

   - The requested odds ratio estimate equals $e^{0.53} = 1.699$.
   - The approx. 95% C.I. for the log-odds-ratio is

   $$0.53 \pm 1.96 \times 0.1166 = (0.3015,\ 0.7585)$$

   - Hence the approx. 95% C.I. for the referred odds ratio is

   $$\left(e^{0.3015}, e^{0.7585}\right) = (1.352,\ 2.135)$$

8. Using the model in 3, estimate the odds ratio of CHD when `BP` increases by two levels and `CHOL` is kept unchanged. Also find an approximately 95% confidence interval for this odds ratio.

   - The requested odds ratio estimate equals $e^{2\times 0.4405} = 2.413$.

- The approx. 95% C.I. for the log-odds-ratio is

$$2 \times 0.4405 \pm 1.96 \times 2 \times 0.1091 = (0.453,\ 1.309)$$

- Hence the approx. 95% C.I. for the referred odds ratio is

$$\left(e^{0.453}, e^{1.309}\right) = (1.574,\ 3.701)$$

9. Using the model in 3, estimate the probability of CHD when `CHOL` is at level 4 and `BP` is at level 3. Also find an approximately 95% confidence interval for this probability.

- At `CHOL=4` and `BP=3`, the estimated log-odds of CHD $= -5.09+4\times0.53+3\times0.44 = -1.65$, with its standard error equal to

$$\sqrt{0.196 + 4^2 \cdot 0.0136 + 3^2 \cdot 0.0119 + 2 \cdot 4 \cdot (-0.0390) + 2 \cdot 3 \cdot (-0.0263) + 2 \cdot 4 \cdot 3 \cdot (-0.00103)}$$

  which equals 0.163.

- The estimated probability $= \dfrac{e^{-1.65}}{1 + e^{-1.65}} = 0.161.$
- An approx. 95% C.I. for the log-odds is $-1.65 \pm 1.96 \times 0.163 = (-1.97,\ -1.33)$. Hence the approx. 95% C.I. for the referred probability is

$$\left(\frac{e^{-1.97}}{1 + e^{-1.97}},\ \frac{e^{-1.33}}{1 + e^{-1.33}}\right) = (0.122,\ 0.209)$$

10. Two people A and B were included in this study. People A had his `CHOL` at level 3 and `BP` at level 1, while people B had his `CHOL` at level 1 and `BP` at level 3. Estimate the odds ratio in regard to CHD for people A versus B based on using the model in 3. Also calculate an approximate 95% confidence interval for this odds ratio.

- The referred log-odds-ratio of CHD for A vs. B equals

$$((1,3,1) - (1,1,3)) \cdot (-5.0916, 0.53, 0.4405)^\top = 0.179$$

  with its standard error equal to

$$\sqrt{(0,2,-2) \cdot \text{summary(fit.5)\$cov.scaled} \cdot (0,2,-2)^\top} = 0.332.$$

- The estimated odds ratio $= e^{0.179} = 1.196.$
- An approx. 95% C.I. for the referred odds ratio

$$\left(e^{0.179-1.96\times0.332},\ e^{0.179+1.96\times0.332}\right) = (0.624,\ 2.293).$$

- Since this C.I. contains 1, there is no significant difference of odds of CHD between A and B at 0.05 significance level.