

Assignment 2 Solutions

1. (1998 Examination, Question 5.) (9 points)

(a) (2 marks) Find the Pearson residuals for each of the cells in the table under the independence model. Present the values in a 2×2 table.

Ans:

	Accept	Reject	
Men	353 (354.2)	207 (205.8)	560
Women	17 (15.8)	8 (9.18)	25
	370	215	585

The Pearson residuals are

	Accept	Reject
Men	-0.06	0.08
Women	0.30	-0.39

(b) (2 marks) Test the hypothesis that being accepted or rejected was independent of the gender of the applicant, using a chi-square statistic.

Ans:

`> 1 - pchisq(0.25, df = 1)`

`[1] 0.6170751`

The chi-square statistics is 0.25. P-value $P(\chi_1^2 \geq 0.25) > 0.05$.

Accept the null hypothesis.

(c) (1 mark) Estimate the log odds ratio of rejection/acceptance of females to males.

Ans: The log-odds ratio $\log \hat{\theta} = \log\left(\frac{353 \times 8}{207 \times 17}\right) = -0.22$.

(d) (2 marks) Calculate the 95% confidence interval for the **odds ratio**. Use the confidence interval argument to verify your conclusion in part(ii). **Ans:** $s.e(\log \hat{\theta}) =$

$$\sqrt{\frac{1}{353} + \frac{1}{207} + \frac{1}{8} + \frac{1}{17}} = 0.44$$

95% CI for log-odds ratio is $-0.22 \pm 1.96 * 0.44 = (-1.08, 0.64)$

95% CI for odds ratio is $(e^{-1.08}, e^{0.64}) = (0.34, 1.90)$

Since the interval contains 1, this verified our conclusion that being accepted or rejected was independent of the gender of the applicant.

(e) (2 marks)

	Successful	Unsuccessful
Males	4	46
Females	1	29

Use fisher exact test because of the low number of successful females applicants.
 The p-value for the test is
 $p = P(\text{no. of successful female} \leq 1 \mid \text{Total number of successful applicants} = 5)$

$$\frac{\binom{50}{4}\binom{30}{1} + \binom{50}{5}\binom{30}{0}}{\binom{80}{5}}$$

$$\frac{690900 + 2118760}{24040016} = 0.38$$

Accept the null hypothesis and conclude that there is NO evidence to support the claim that successful scholarship application is lower in females.

2. (4 points)

Consider a 3-way contingency table where we observe three response variables R, S and T which have r, s and t categories respectively. Let

$$p_{ijk} = P(R = i, S = j, T = k), \quad i = 1, \dots, r; j = 1, \dots, s; k = 1, \dots, t.$$

R to S

S is independent of T given R means

$$\frac{p_{ijk}}{p_{.j\cdot}} = P(S = j, T = k \mid R = i) = P(S = j \mid R = i)P(T = k \mid R = i).$$

That is $p_{ijk} = \frac{p_{ij\cdot}p_{i\cdot k}}{p_{i\cdot\cdot}}$.

From lectures $(\alpha\beta\gamma)_{ijk} = (\beta\gamma)_{jk} = 0$ implies $p_{ijk}p_{i11} = p_{ij1}p_{i1k}$ (1)
 as

$$(\beta\gamma)_{jk} = \log \left(\frac{p_{ijk}p_{i11}}{p_{ij1}p_{i1k}} \right).$$

Sum (1) over j gives $p_{i\cdot k}p_{i11} = p_{i\cdot 1}p_{i1k}$ (2)

Sum (1) over k gives $p_{ij\cdot}p_{i11} = p_{ij1}p_{i1\cdot}$ (3)

Sum (2) over k gives $p_{i\cdot\cdot}p_{i11} = p_{i\cdot 1}p_{i1\cdot}$ (4)

(4) yields

$$p_{i11} = \frac{p_{i\cdot 1}p_{i1\cdot}}{p_{i\cdot\cdot}}$$

as required for $i = k = 1$.

Substitute (2), (3) and (4) into (1)

$$p_{ijk}p_{i11} = \frac{p_{ij} \cdot p_{i11}}{p_{i1\cdot}} \times \frac{p_{i\cdot k} p_{i11}}{p_{i\cdot 1}}$$

$$p_{ijk} = \frac{p_{ij} \cdot p_{i\cdot k}}{p_{i\cdot \cdot}}$$

as $p_{i1\cdot} p_{i\cdot 1} = p_{i\cdot \cdot} p_{i11}$.

3. (9 points)

(a) (2 points)

Test the hypothesis of complete independence of the three variables using the deviance.

```
> y <- c(716, 79, 207, 25, 819, 67, 186, 22)
> p <- factor(c(1, 1, 1, 1, 2, 2, 2, 2))
> bp <- factor(c(1, 2, 1, 2, 1, 2, 1, 2))
> cc <- factor(c(1, 1, 2, 2, 1, 1, 2, 2))
> glm1 <- glm(y ~ p + bp + cc, family = poisson)
> summary(glm1)
```

Call:

```
glm(formula = y ~ p + bp + cc, family = poisson)
```

Deviance Residuals:

	1	2	3	4	5	6	7	8
	-0.8829	0.5672	0.9476	1.2198	1.0917	-1.3754	-1.4374	0.2937

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	6.60649	0.03383	195.298	<2e-16 ***
p2	0.06320	0.04345	1.455	0.146
bp2	-2.30155	0.07550	-30.485	<2e-16 ***
cc2	-1.34037	0.05355	-25.030	<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: 2432.612 on 7 degrees of freedom
Residual deviance: 8.723 on 4 degrees of freedom
AIC: 70.171
```

Number of Fisher Scoring iterations: 4

```
> 1 - pchisq(8.7229, 4)
```

```
[1] 0.06841142
```

The data are presented in a $2 \times 2 \times 2$ table. to fit the model of complete independence for personality, diastolic blood pressure and cholesterol we fit a log-linear model with no interaction terms. The deviance for this model (from the R output) is $D = 8.723$ with 4 d.f. $p\text{-value} = P(\chi_4^2 \geq 8.723) = 0.068$.

Thus we accept the null and conclude that the model gives reasonable fit.

(b) (2 points)

```
> glm1$fitted
      1      2      3      4      5      6      7      8
739.88436 74.06519 193.66396 19.38649 788.15335 78.89709 206.29832 20.65123

> r <- (y - glm1$fitted)/sqrt(glm1$fitted)
> matrix(r, ncol = 2, byrow = T)
      [,1]      [,2]
[1,] -0.8780753  0.5734077
[2,]  0.9583020  1.2749269
[3,]  1.0987594 -1.3394003
[4,] -1.4132280  0.2968002
```

The Pearson residual for the (i, j, k) cell is

$$r_{ijk} = \frac{O_{ijk} - e_{ijk}}{\sqrt{e_{ijk}}}.$$

The table of residuals is

Personality(P)	Cholesterol (C)	Diastolic	B.P. (D)
		Normal	High
A	Normal	-0.878	0.573
	High	0.958	1.275
B	Normal	1.099	-1.339
	High	-1.413	0.297

All residuals are small, less than 1.5 in absolute value, consistent with the independence model giving reasonable fit. Test the fit of this model using Pearson's χ^2 is

$$\chi^2 = \sum_i \sum_j \sum_k \left(\frac{O_{ijk}^2}{e_{ijk}} \right) - 2121 = 2129.73 - 2121 = 8.73$$

, which is same as part (a).

```
> o = y
> e = glm1$fitted
> sum(o^2/e)
[1] 2129.73
> sum(o)
[1] 2121
```

(c) (2 points)

Compare cholesterol level for personality types A and B.

```
> a <- c(716 + 79, 207 + 25, 819 + 67, 186 + 22)
```

```
> am <- matrix(a, ncol = 2, byrow = T)
```

```
> am
```

```
      [,1] [,2]
[1,]  795  232
[2,]  886  208
```

```
> chisq.test(am, correct = FALSE)
```

Pearson's Chi-squared test

```
data: am
```

```
X-squared = 4.123, df = 1, p-value = 0.04230
```

The marginal table for comparing personality and cholesterol levels is provided the matrix `am`.

Estimated proportion of type A personalities with high cholesterol levels is $\hat{p}_A = \frac{232}{232+795} = 0.226$

The corresponding estimate for type B personalities is $\hat{p}_B = \frac{208}{208+886} = 0.190$

Testing for independence of personality type and blood pressure in the marginal table gave $\chi^2 = 4.12$ with p-value $P(\chi_1^2 \geq 4.12) = 0.0423$.

So again, we reject that personality type and cholesterol levels are independent.

(d) (2 points)

The marginal table comparing diastolic blood pressure for each personality type is

	Normal	High
A	923	104
B	1005	89

Estimated log-odds ratio is

$$\log(\hat{\theta}) = \log \frac{923 \times 89}{104 \times 1005} = \log(0.786) = -0.241.$$

An approximate 95 % C.I. for $\log(\hat{\theta})$ is

$$\begin{aligned} & -0.241 \pm 1.96 \sqrt{\frac{1}{923} + \frac{1}{104} + \frac{1}{1005} + \frac{1}{89}} \\ & -0.241 \pm 0.297 \end{aligned}$$

The C.I. = $(-0.538, 0.056)$.

(e) (1 point)

```
> glm1$coef
```

```

(Intercept)          p2          bp2          cc2
 6.60649391  0.06319877 -2.30154829 -1.34036941
> glm1$fit
      1          2          3          4          5          6          7          8
739.88436  74.06519 193.66396  19.38649 788.15335  78.89709 206.29832 20.65123
> glm1$fit[4]/sum(glm1$fit[1:4])
      4
0.01887681

```

Estimate the proportion of Type A personality with high blood pressure and high cholesterol level is 0.01887 (1.89%).

4. (6 points)

(a) (2 points)

```

> n <- c(59, 60, 62, 56, 63, 59, 62, 60)
> y <- c(6, 13, 18, 28, 52, 53, 61, 60)
> d <- c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.861, 1.8939)
> yn <- cbind(y, (n - y))
> glm1 <- glm(yn ~ d, family = binomial)
> summary(glm1)

```

Call:

```
glm(formula = yn ~ d, family = binomial)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5657	-0.3582	0.8448	1.2619	1.3686

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-60.282	5.184	-11.63	<2e-16 ***
d	34.022	2.915	11.67	<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: 10.491 on 6 degrees of freedom
AIC: 40.689

Number of Fisher Scoring iterations: 4

```
> glm1$fitted
```

```

      1      2      3      4      5      6      7
0.05946507 0.16502137 0.36200856 0.60347695 0.79281257 0.90140372 0.95403604
      8
0.98451320

```

```
> plot(d, glm1$fitted, main = "Logistic fit to beetle data")
```

The fitted logistic regression model is:

$$\log\left(\frac{p}{1-p}\right) = -60.282 + 34.022x$$

where $1.69 < x < 1.89$.

- (b) (2 points) Calculate an approximate 90% confidence interval for the coefficient of x in the logistic regression model.

An approximate 90% C.I. for the coefficient of x is

$$34.022 \pm 1.64 \times \text{estimated s.e.}$$

i.e.

$$34.022 \pm 1.64 \times 2.915 = 34.022 \pm 4.78$$

$$(29.2, 38.8).$$

- (c) (1 point) What (log) dose results in a 20% success rate? If $p = 0.2$ then $\log\left(\frac{p}{1-p}\right) = \log\frac{0.2}{0.8} = -1.386$. So the estimated log dose for a 20% success rate is

$$\frac{-1.386 + 60.282}{34.022} = 1.73.$$

- (d) (1 point)

$$\hat{p} = \frac{e^{-60.282+34.022 \times 1.5}}{1 + e^{-60.282+34.022 \times 1.5}} = \frac{e^{-9.25}}{1 + e^{-9.25}} \approx 0$$

Predict the number of beetles surviving is $70(1 - \hat{p}) = 70$. That is, predict 70 will survive none will die.

5. (2 points)

	Successful	Unsuccessful
Males	4	76
Females	1	49

Use fisher exact test because of the low number of successful females applicants. The p-value for the test is

$p = P(\text{no. of successful female} \leq 1 \mid \text{Total number of successful applicants} = 5)$

$$\frac{\binom{80}{4}\binom{50}{1} + \binom{80}{5}\binom{50}{0}}{\binom{130}{5}}$$

$$\frac{79079000 + 24040016}{286243776} = 0.3602$$

Accept the null hypothesis and conclude that there is NO evidence to support the claim that successful scholarship application is lower in females.