

MATH 523 - ASSIGNMENT 2: SOLUTIONS

(a) Five basic models can be fitted:

M0 Null
M1 age
M2 test
M3 age+test
M4 age+test+age.test

Here model M4 is the saturated model, the most complicated we can fit for this two factor model. The R code for fitting these models is given below:

```
1 y<-c(22,26,14,30,32,23,29)
2 z<-c(52,55,59,53,31,28,17)
3 m<-y+z
4
5 Y.p<-c(y,z)
6 Age<-gl(7,1,14)
7 Test<-gl(2,7)
8
9 glm0<-glm(Y.p~1,family=poisson)
10 glm1<-glm(Y.p~Age,family=poisson)
11 glm2<-glm(Y.p~Test,family=poisson)
12 glm3<-glm(Y.p~Age+Test,family=poisson)
13 glm4<-glm(Y.p~Age*Test,family=poisson)
14
15 anova(glm4)
16 summary(glm4)
```

The result of the anova call on line 18 is the following

```
> anova(glm4)
Analysis of Deviance Table
```

Model: poisson, link: log

Response: Y.p

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			13	82.427
Age	6	19.324	7	63.103
Test	1	30.394	6	32.709
Age:Test	6	32.709	0	0.000

This allows the comparison of all models; it is clear that the difference in deviance when simplifying from M4 to M3 (32.709 on 6 degrees of freedom) is significant (we have also that $\chi^2_6(0.95) = 12.592$), and that M3 is **not** an adequate simplification of M4. Therefore, in terms of model adequacy, we are left with the saturated model

$$\text{age} + \text{test} + \text{age.test} \equiv \text{age} * \text{test}$$

Clearly this model fits the data exactly, and has deviance zero by definition.

The result of the summary call on line 19 is the following

```
> summary(glm4)

Call: glm(formula = Y.p ~ Age * Test, family = poisson)

Deviance Residuals:
 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.09104    0.21320  14.498 < 2e-16 ***
Age2          0.16705    0.28968   0.577 0.564156
Age3         -0.45199    0.34188  -1.322 0.186151
Age4          0.31015    0.28069   1.105 0.269174
Age5          0.37469    0.27696   1.353 0.176088
Age6          0.04445    0.29822   0.149 0.881507
Age7          0.27625    0.28273   0.977 0.328527
Test2         0.86020    0.25433   3.382 0.000719 ***
Age2:Test2   -0.11096    0.34832  -0.319 0.750054
Age3:Test2    0.57828    0.39123   1.478 0.139383
Age4:Test2   -0.29111    0.34189  -0.851 0.394508
Age5:Test2   -0.89195    0.35804  -2.491 0.012731 *
Age6:Test2   -0.66349    0.37931  -1.749 0.080257 .
Age7:Test2   -1.39428    0.39748  -3.508 0.000452 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 8.2427e+01  on 13  degrees of freedom
Residual deviance: 2.6645e-15  on  0  degrees of freedom
AIC: 101.78

Number of Fisher Scoring iterations: 3
```

Remember that R reports the contrasts from baseline on the linear predictor scale, that is, here for row $j = 1, 2$, column $k = 1, \dots, 7$.

$$\log \mu_{jk} = \begin{cases} \beta_1 & j = 1, k = 1 \\ \beta_1 + \beta_{1k} & j = 1, k = 2, \dots, 7 \\ \beta_1 + \beta_{21} & j = 2, k = 1 \\ \beta_1 + \beta_{1k} + \beta_{21} + \gamma_{2k} & j = 2, k = 2, \dots, 7 \end{cases}$$

so that, for example, $\hat{\beta}_1 = 3.09104$, $\hat{\beta}_{12} = 0.16705$, $\hat{\beta}_{21} = 0.86020$, $\hat{\gamma}_{22} = -0.11096$ etc. For example, the fitted value for the $j = 1, k = 5$ cell is

$$\log \hat{\mu}_{15} = \hat{\beta}_1 + \hat{\beta}_{15} = 3.09104 + 0.37469 = 3.46573$$

which on exponentiation yields $\hat{y}_{15} = \mu_{15} = 32 = y_{15}$.

The fitted value for the $j = 2, k = 3$ cell is

$$\log \hat{\mu}_{23} = \hat{\beta}_1 + \hat{\beta}_{13} + \hat{\beta}_{21} + \hat{\beta}_{23} = 3.09104 + (-0.45199) + 0.86020 + 0.57828 = 4.07753$$

which on exponentiation yields $\hat{y}_{23} = \mu_{23} = 59 = y_{23}$.

8 MARKS

- (b) For the binomial version of the analysis, **conditional on column totals**, there are now only two models to fit:

M0 Null

M1 age

The R code for fitting the models using the canonical logistic link is

```
1 Age.b<-gl(7,1)
2
3 glm0.b<-glm(cbind(y,z)~1,family=binomial)
4 glm1.b<-glm(cbind(y,z)~Age.b,family=binomial)
5 summary(glm1.b)

> summary(glm1.b)

Call: glm(formula = cbind(y, z) ~ Age.b, family = binomial)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8602     0.2543  -3.382  0.000719 ***
Age.b2         0.1110     0.3483   0.319  0.750054
Age.b3        -0.5783     0.3912  -1.478  0.139383
Age.b4         0.2911     0.3419   0.851  0.394508
Age.b5         0.8919     0.3580   2.491  0.012731 *
Age.b6         0.6635     0.3793   1.749  0.080257 .
Age.b7         1.3943     0.3975   3.508  0.000452 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  3.2709e+01  on 6  degrees of freedom
Residual deviance: -2.4647e-14  on 0  degrees of freedom
AIC: 45.582

Number of Fisher Scoring iterations: 3
```

Again, M1 is a saturated model; we have seven data points (reduced from 14 in the Poisson analysis by conditioning on the seven column totals), and seven parameters. R reports the contrasts from baseline on the linear predictor scale, that is, here for column $k = 1, \dots, 7$.

$$\log \left(\frac{\pi_k}{1 - \pi_k} \right) = \begin{cases} \theta_1 & k = 1 \\ \theta_1 + \theta_k & k = 2, \dots, 7 \end{cases}$$

The estimate of θ_1 is -0.8602, and the estimates of $\theta_k, k = 2, \dots, 7$ are contained in the subsequent rows of the coefficients table. Remember that

$$\theta_k = \log \left(\frac{\pi_k}{1 - \pi_k} \right) - \log \left(\frac{\pi_1}{1 - \pi_1} \right) = \log \left(\frac{\pi_k / (1 - \pi_k)}{\pi_1 / (1 - \pi_1)} \right)$$

so that θ_k is the log odds ratio comparing column k to column 1.

It is clear again that this model fits the data exactly, yielding a deviance of zero.

6 MARKS

- (c) The model M4 for the Poisson case, and model M1 for the Binomial case, both fit the data exactly, albeit under different assumptions. This suggests that there must be some underlying connection between them; the connection is given by the following argument. Consider the Poisson variates, presumed independent, in the first analysis, $Y_{jk}, j = 1, 2, k = 1, \dots, 7$. We have, of course, that $Y_{jk} \sim \text{Poisson}(\mu_{jk})$. Now consider the k th column total, $M_k, k = 1, \dots, 7$. We have, by basic results concerning the Poisson distribution, that

$$M_k \sim \text{Poisson}(\mu_{1k} + \mu_{2k}) \quad k = 1, \dots, 7.$$

Finally, consider the conditional distribution of $Y_k \equiv Y_{1k}$, given that $M_k = m_k$: by elementary calculations for discrete distributions, we have that

$$Y_k | M_k = m_k \sim \text{Binomial}(m_k, \pi_k)$$

where $\pi_k = \mu_{1k} / (\mu_{1k} + \mu_{2k})$. This must coincide with the model fitted to the Binomial data in part (b), and suggests that, in the fits of the two models, we can connect the parameter estimates using this identity, that is, that

$$\hat{\pi}_k = \frac{\hat{\mu}_{1k}}{\hat{\mu}_{1k} + \hat{\mu}_{2k}}.$$

This turns out to be the case; the ML estimates of $\pi_1, \dots, \pi_7$ are simply the column proportions

$$\frac{y_k}{m_k} = \frac{y_k}{y_k + z_k} \equiv \frac{y_{1k}}{y_{1k} + y_{2k}}$$

and so the estimates match precisely after transformation.

4 MARKS

- (d) The pattern of results across age categories (depicted below), suggests focussing on an **increasing** - perhaps **linear** - trend in the proportions of positive tests across age groups.

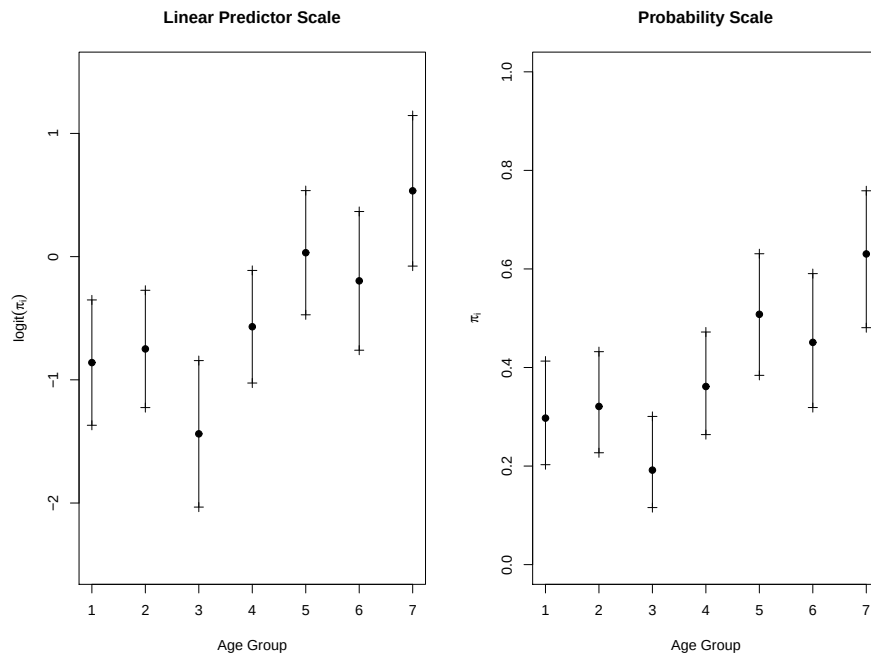


Figure 1: Fitted linear predictor and probability values across age groups

2 MARKS