R Code And Tasks Chapter 6 (MAS 6003)

Witold Wolski

December 28, 2016

6.1 Types of 2 way tables - response / control variables

6.1.1 Case(a): Skin cancer (melanoma) data - 2 response variables

Cross sectional study of malignant melanoma. Both tumor type and site are **response variables** because none of the row or column totals were fixed in advance of the data collection.

```
library(reshape2)
rm(list=ls())
load("data/MAS367-GLMs.RData", envir = e <- new.env())

Mela <- e$Mela
head(Mela)
```

```
##   number tumour.type          site
## 1     22           A Head.and.Neck
## 2     16           B Head.and.Neck
## 3     19           C Head.and.Neck
## 4     11           D Head.and.Neck
## 5      2           A         Trunk
## 6     54           B         Trunk
```

```
dcast(Mela,tumour.type~ site, value.var="number")
```

```
##   tumour.type Head.and.Neck Trunk Extremities
## 1           A            22     2          10
## 2           B            16    54         115
## 3           C            19    33          73
## 4           D            11    17          28
```

6.1.2 Case(b) : Flu vaccine data - 1 response and 1 control variable

Patients were randomly assigned to the two groups (Placebo, Vaccine), and the response (levels of an antibody found in the blood six weeks after vaccination) was determined. Antibody level is the **response** and vaccine group is a **controlled variable** (with totals fixed by experimental design).

```
vaccine <- e$vaccine
head(vaccine)
```

```
##   count response treatment
## 1    25    small   placebo
## 2     6    small   vaccine
## 3     8 moderate   placebo
## 4    18 moderate   vaccine
## 5     5    large   placebo
## 6    11    large   vaccine
```

```
dcast(vaccine, treatment ~ response, value.var="count")
```

```
##   treatment small moderate large
## 1   placebo    25        8     5
## 2   vaccine     6       18    11
```

6.2.1 Association, Independence and Homogeneity

Independence :

Case(a) - skin cancer

Probabilities of interest $\pi_{ij} = P(A = i, B = j)$

$$P(A = i, B = j) = P(A = i) \times P(B = j)$$

for all i and j, where $\pi_i = P(A = i)$ and $\pi_j = P(B = j)$ are the marginal probabilities of row i and column j

Case(b) - flue vaccine data

Probabilities of interest are conditional probabilities $\pi ij = P(B = j|A = i)$.

The interest is in whether the probability distribution of the response (antibody level) is the same in each level of the controlled variable (drug group). If it doesn't depend on i then we can write

$$(\pi_{ij} = \pi_{.j})$$

.

This is known as **homogeneity**.

6.3 Distribution for two-way tables

6.3.1 Case(a): two response variables

6.3.2 Case(b): one response variable

6.3.3 Case(c): independent poisson (no fixed margins).

6.3.4 Expected values

6.4 GLMs and two-way contingency tables

6.4.1 Natural hypothesis are log-linear models

6.4.2 Poisson log-linear modelling for two-way tables

6.4.3 Maximum likelihood estimation for π_{ij} in case (a)

Task 19

Verify the maximum likelihood estimate for π_{ij} for the A + B model for case(a).
Verify that $\pi_{ij} = \frac{y_{i\cdot} y_{\cdot j}}{n^2}$
and somehow use information that $\mu_{ij} = n\pi_i\pi_j$

Task 20

Verify the maximum likelihood estimate for π_{ij} for the A + B model for case(b).
Verify that $\pi_{ij} = \frac{y_{\cdot j}}{n}$
and somehow use information that $\mu_{ij} = n_i\pi_j$

$$\begin{aligned}
\partial l(\mu_i)/\partial\pi_j &= (\sum_i\sum_j -n_i\pi_j + y_{ij}\log(n_i\pi_j) - \log(y_{ij}!))' \\
&= \sum_i -n_i + \frac{y_{ij}}{\pi_j} \\
&= n + \frac{y_{ij}}{\pi_j} = 0 \\
\pi_{ij} &= \frac{y_{\cdot j}}{n}
\end{aligned}$$

6.5 Interaction plots and examination of residues.

`interaction.plot` honorable mentioned.

6.6 Analysis of the skin cancer data (case(a)) using log-linear models

Task 21

Calculate the table of fitted values for the linear predictor containing B for case(a)

6.6.1 Fitted values for the skin cancer data

- $A \times B$ saturated model.

```
glm.sat <- glm(number ~ factor(tumour.type) * factor(site), family = poisson(log), data=Mela)
matrix(glm.sat$fitted.values,ncol=3)
```

```
##      [,1] [,2] [,3]
## [1,]   22    2   10
## [2,]   16   54  115
## [3,]   19   33   73
## [4,]   11   17   28
```

- $A + B$:independence

```
glm.indep <- glm(number ~ factor(tumour.type) + factor(site), family = poisson(log), data=Mela)
glm.indep
```

```
## 
## Call:  glm(formula = number ~ factor(tumour.type) + factor(site), family = poisson(log), 
##     data = Mela)
## 
## Coefficients:
##             (Intercept)     factor(tumour.type)B     factor(tumour.type)C  
##                  1.7544                   1.6940                   1.3020  
##    factor(tumour.type)D        factor(site)Trunk  factor(site)Extremities  
##                  0.4990                   0.4439                   1.2010  
## 
## Degrees of Freedom: 11 Total (i.e. Null);  6 Residual
## Null Deviance:       295.2 
## Residual Deviance: 51.8  AIC: 122.9
```

```
matrix(glm.indep$fitted.values,ncol=3)
```

```
##       [,1]   [,2]    [,3]
## [1,]  5.78  9.010  19.210
## [2,] 31.45 49.025 104.525
## [3,] 21.25 33.125  70.625
## [4,]  9.52 14.840  31.640
```

Or much simpler

```
mel <- aggregate(number ~ site, Mela, sum)$number
tum <- aggregate(number ~ tumour.type, Mela, sum)$number
(mel %*% t(tum))/400
```

```
##       [,1]    [,2]   [,3]  [,4]
## [1,]  5.78  31.450 21.250  9.52
## [2,]  9.01  49.025 33.125 14.84
## [3,] 19.21 104.525 70.625 31.64
```

- A : independence and the same probability for each column category

```
glm.indepCol <- glm(number ~ factor(tumour.type), family = poisson(log), data=Mela)
t(x <- matrix(glm.indepCol$fitted.values,nrow=3, byrow=T))
```

```
##           [,1]     [,2]     [,3]
## [1,] 11.33333 11.33333 11.33333
## [2,] 61.66667 61.66667 61.66667
## [3,] 41.66667 41.66667 41.66667
## [4,] 18.66667 18.66667 18.66667
```

Or much simpler

```
tum / 3
```

```
## [1] 11.33333 61.66667 41.66667 18.66667
```

- B : independence and the same probability for each row category

Calculate the table of fitted values for the linear predictor containing B for case(a)

```
glm.indepRow <- glm(number ~ factor(site), family = poisson(log), data=Mela)
(res<-matrix(glm.indepRow$fitted.values,ncol=3))
```

```
##      [,1] [,2] [,3]
## [1,]   17 26.5 56.5
## [2,]   17 26.5 56.5
## [3,]   17 26.5 56.5
## [4,]   17 26.5 56.5
```

6.6.2

- $A \times B$: saturated model

```
(void <- glm(number ~ factor(tumour.type) * factor(site), family = poisson(log), data=Mela))
```

```
## 
## Call:  glm(formula = number ~ factor(tumour.type) * factor(site), family = poisson(log), 
##     data = Mela)
## 
## Coefficients:
##                                    (Intercept)  
##                                         3.0910  
##                           factor(tumour.type)B  
##                                        -0.3185  
##                           factor(tumour.type)C  
##                                        -0.1466  
##                           factor(tumour.type)D  
##                                        -0.6931  
##                              factor(site)Trunk  
##                                        -2.3979  
##                        factor(site)Extremities  
##                                        -0.7885  
##         factor(tumour.type)B:factor(site)Trunk  
##                                         3.6143  
##         factor(tumour.type)C:factor(site)Trunk  
##                                       2.9500
##       factor(tumour.type)D:factor(site)Trunk
##                                       2.8332
## factor(tumour.type)B:factor(site)Extremities
##                                       2.7608
## factor(tumour.type)C:factor(site)Extremities
##                                       2.1345
## factor(tumour.type)D:factor(site)Extremities
##                                       1.7228
##
## Degrees of Freedom: 11 Total (i.e. Null);  0 Residual
## Null Deviance:       295.2
## Residual Deviance: 8.66e-15  AIC: 83.11
```

```
matrix(void$fitted.values, byrow = F, nrow=4)
```

```
##      [,1] [,2] [,3]
## [1,]   22    2   10
## [2,]   16   54  115
## [3,]   19   33   73
## [4,]   11   17   28
```

6.6.3 Skin cancer data case(a) revisited

1. The test of independence based on the log-linear model $A + B$

```
1 - pchisq(glm.indep$deviance,6)
```

```
## [1] 2.050453e-09
```

```
qchisq(0.975,6)
```

```
## [1] 14.44938
```

2. The usual Pearson χ^2 test for independence

```
chisq.test(matrix(Mela$number,nrow=4))
```

```
##
##  Pearson's Chi-squared test
##
## data:  matrix(Mela$number, nrow = 4)
## X-squared = 65.813, df = 6, p-value = 2.943e-12
```

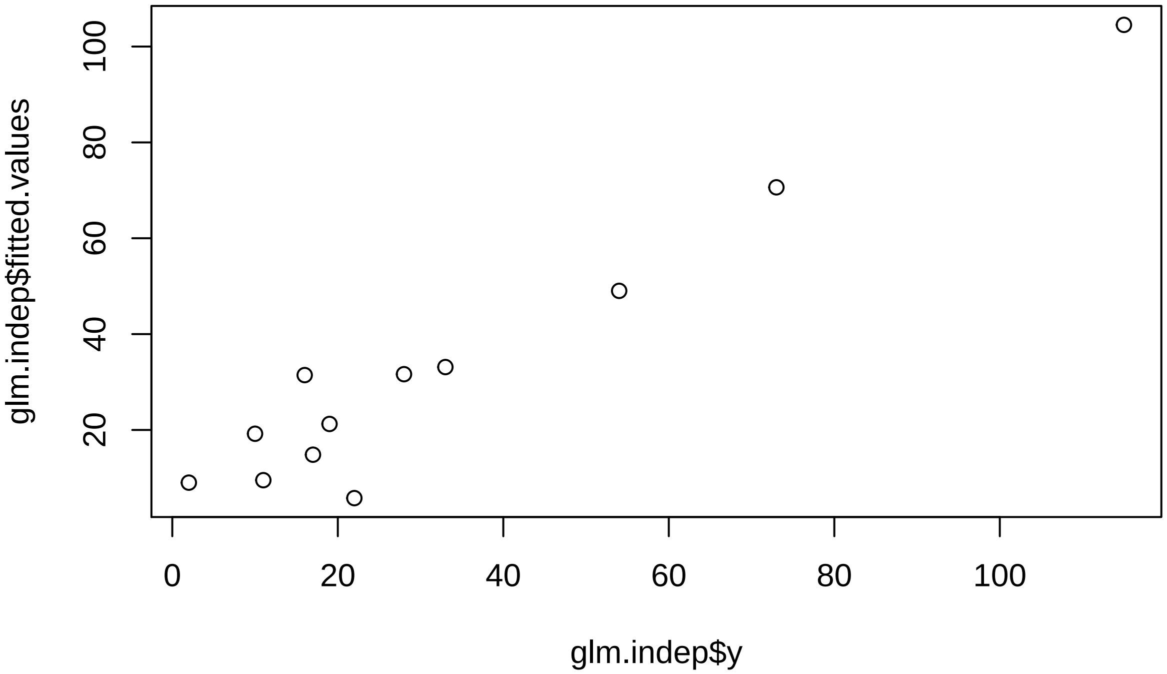
```
1-pchisq(65.813,6)
```

```
## [1] 2.94309e-12
```

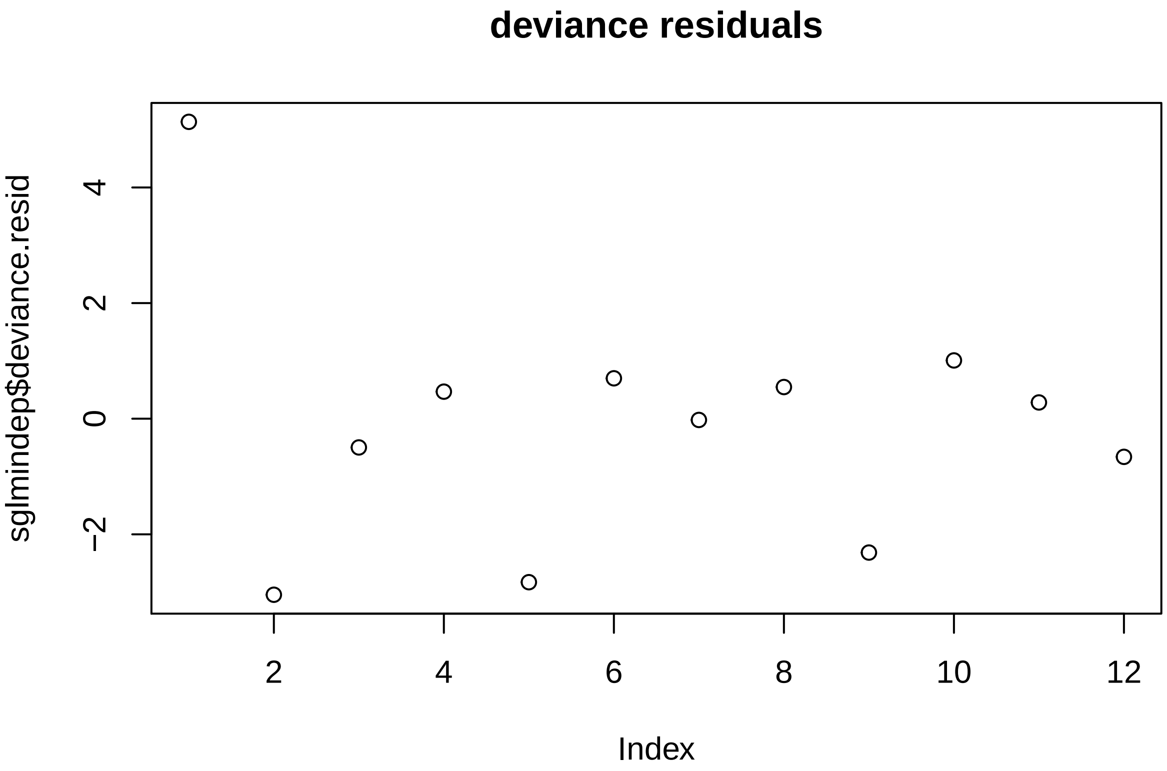
5.

Deviance residuals see 3.6.4 in the notes. Pearson residuals see Task 9 and page 91.

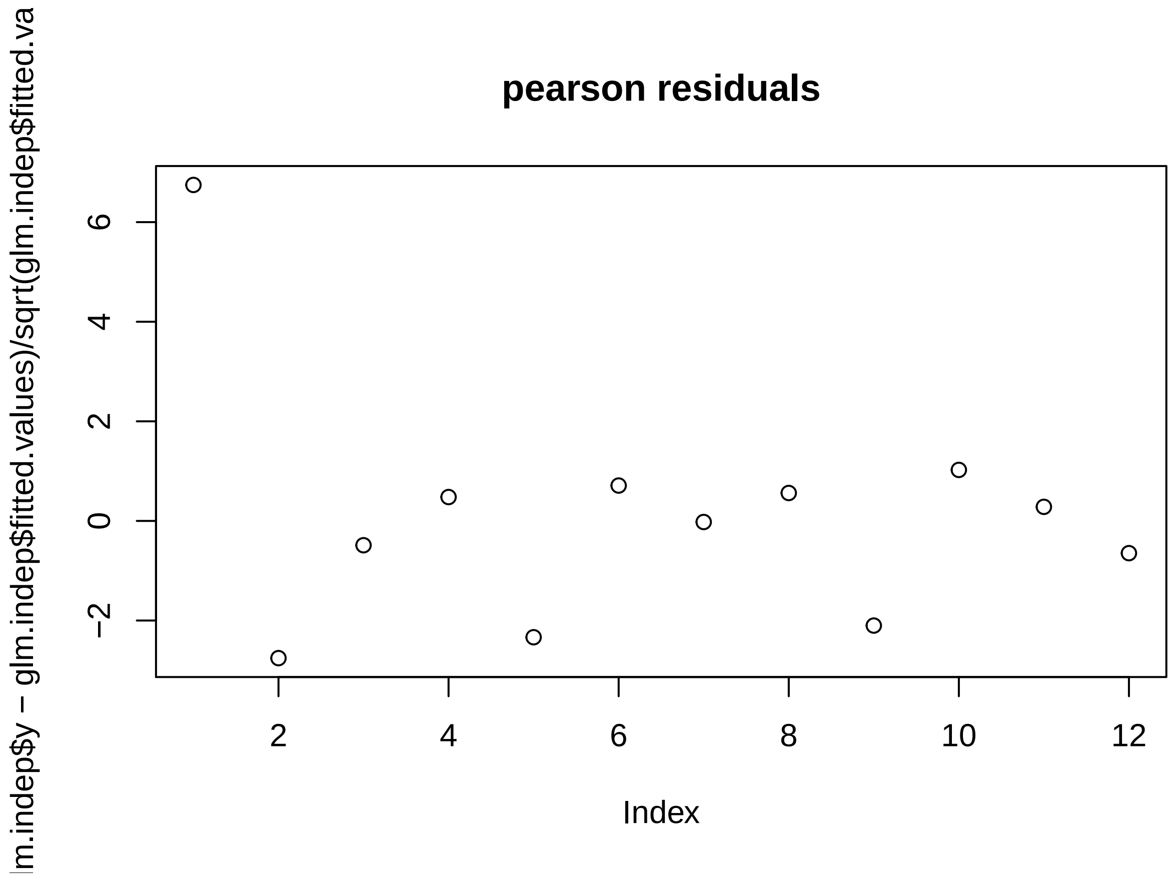
```
plot(glm.indep$y,glm.indep$fitted.values)
```



```
sglmindep<-summary(glm.indep)
plot(sglmindep$deviance.resid , main="deviance residuals")
```



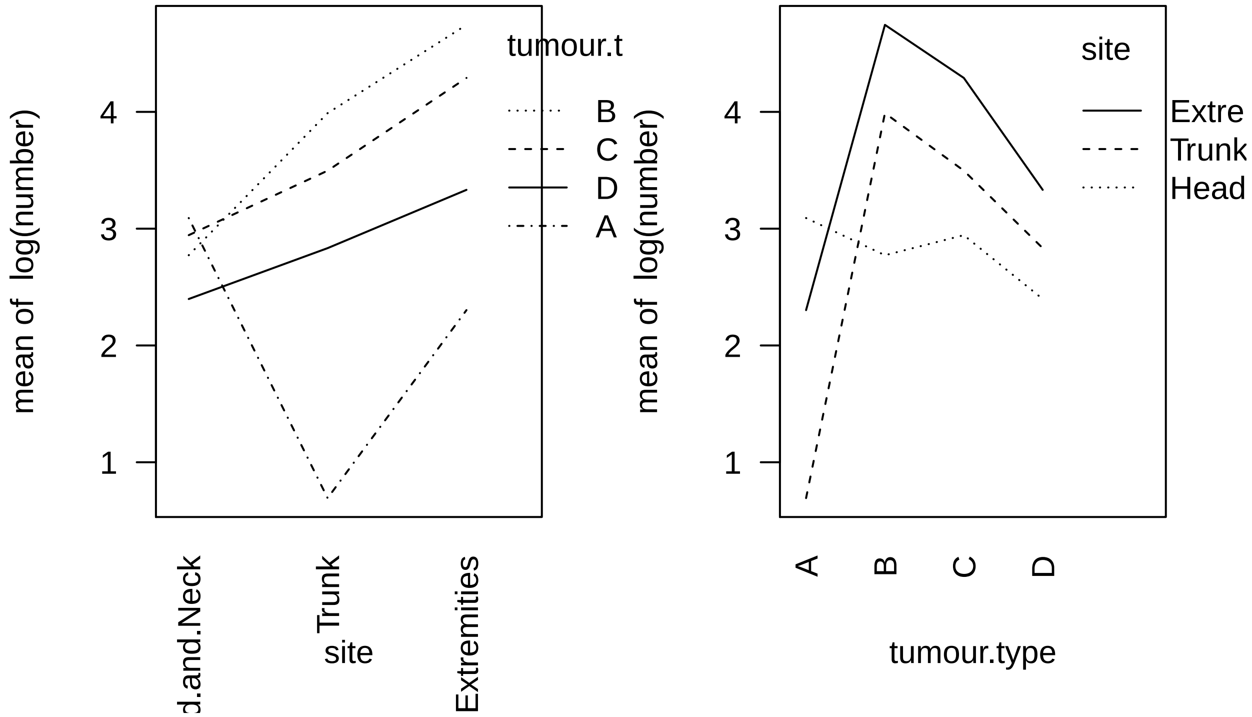
```r
plot((glm.indep$y-glm.indep$fitted.values)/sqrt(glm.indep$fitted.values), main = "pearson residuals")
```



6. interaction plot

```
par(mfrow=c(1,2))
with(Mela, {
interaction.plot(site, tumour.type, log(number),las=2)
interaction.plot(tumour.type, site, log(number),las=2)

})
```



8. removing first row (T1) or first column (Head and Neck)

- remove cancer type A

```
MelaNoA <- Mela[Mela$tumour.type!="A",]
glm.indepNoA <- glm(number ~ factor(tumour.type) + factor(site), family = poisson(log), data=MelaNoA)
glm.indepNoA
```

```
##
## Call:  glm(formula = number ~ factor(tumour.type) + factor(site), family = poisson(log),
##     data = MelaNoA)
##
## Coefficients:
##             (Intercept)     factor(tumour.type)C     factor(tumour.type)D
##                  3.1464                  -0.3920                  -1.1950
##       factor(site)Trunk  factor(site)Extremities
##                  0.8157                   1.5466
##
## Degrees of Freedom: 8 Total (i.e. Null);  4 Residual
## Null Deviance:       203.3
## Residual Deviance: 6.509     AIC: 63.91
```

```
qchisq(0.975, 4)
```

```
## [1] 11.14329
```

- remove site Head

```
MelaNoHead <- Mela[Mela$site!="Head.and.Neck",]

glm.indepNoHead <- glm(number ~ factor(tumour.type) + factor(site), family = poisson(log), data=MelaNoH
glm.indepNoHead
```

```
##
## Call:  glm(formula = number ~ factor(tumour.type) + factor(site), family = poisson(log),
##     data = MelaNoHead)
##
## Coefficients:
##            (Intercept)     factor(tumour.type)B     factor(tumour.type)C
##                 1.3432                   2.6450                   2.1785
##   factor(tumour.type)D  factor(site)Extremities
##                 1.3218                   0.7571
##
## Degrees of Freedom: 7 Total (i.e. Null);  3 Residual
## Null Deviance:       237.2
## Residual Deviance: 2.165     AIC: 52.68
```

```
qchisq(0.975, 3)
```

```
## [1] 9.348404
```

9. and 10. is about pooling the B, C, D or Trunk and Extremities

10.

A log-linear model can be fitted which is additive in the factors, but includes a term for the (1,1) cell — an indicator variable for that cell (that is, treats it as an outlier).

```
head(Mela)
```

```
##   number tumour.type          site
## 1     22           A Head.and.Neck
## 2     16           B Head.and.Neck
## 3     19           C Head.and.Neck
## 4     11           D Head.and.Neck
## 5      2           A         Trunk
## 6     54           B         Trunk
```

```
MelaFix <- Mela
MelaFix$labelAHead <- rep(0,nrow(Mela))
MelaFix$labelAHead[1] <- 1
MelaFix$labelAHead <- as.factor(MelaFix$labelAHead)

glm.indepNoHead <- glm(number ~ factor(tumour.type) + factor(site) + factor(labelAHead), family = poiss
glm.indepNoHead
```

```
##
## Call:  glm(formula = number ~ factor(tumour.type) + factor(site) + factor(labelAHead),
##     family = poisson(log), data = MelaFix)
##
## Coefficients:
##            (Intercept)     factor(tumour.type)B     factor(tumour.type)C
##                 0.5452                   2.6011                   2.2091
##   factor(tumour.type)D        factor(site)Trunk  factor(site)Extremities
##                 1.4061                   0.7980                   1.5551
##    factor(labelAHead)1
##                   2.5458
##
## Degrees of Freedom: 11 Total (i.e. Null);  5 Residual
## Null Deviance:       295.2
## Residual Deviance: 8.002     AIC: 81.11
```

Task 22 Verify the analysis (See above)

Task 23

For the 4×3 table in Example 6.1.1, and the independence model, show directly (with- out fitting a log-linear model) that $\mu_{11} = 5.780$, $e_{P,11} = 6.747$, and $e_{D,11} = 5.135$.

$e_{P,11}$ - Pearson resdiual

$$\begin{aligned} e_{P,i} &= w_i \frac{y_i - \hat{y}_i}{\sqrt{V(\hat{\mu_i})}} \\ &= \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}} = \frac{22 - 5.780}{\sqrt{5.780}} \end{aligned}$$

For poisson distribution. For other distributions, see Taks 9.

$e_{D,11}$ - Deviance residual.

The i'th deviance residual is defined by

$$e_{D,i} = sign(y_i - \hat{y}_i) \times \sqrt{d_i}$$

$$d_i = 2 \times \{y_i \log(\frac{y_i}{\mu_i}) - (y_i - \mu_i)\}$$

```r
y <- 22
mu <- 5.78
sign(y-mu) * (sqrt(2* (y*log(y/mu) - (y-mu))))
```

```
## [1] 5.135378
```

Task 24

Consider the $2x3$ two-way table $\{y_{ij}, i = 1, 2, j = 1, 2, 3\}$, with factors A, B. Let the deviance for the additive model $A + B$ be D for the 2Ö3 table, D_1 be for the $2x2$ sub-table $\{y_{ij}, i = 1, 2, j = 1, 2\}$, and D_2 for the $2x2$ table $\{z_{ij}, i = 1, 2, j = 1, 2\}$, where $z_{i1} = y_{i1} + y_{i2}, z_{i2} = y_{i3}$ for $i = 1, 2$. Verify that $D = D_1 + D_2$.

6.7 Flu vaccine data (case(b)) revisited

- minimal model

```r
head(vaccine)
```

```
##   count response treatment
## 1    25    small   placebo
## 2     6    small   vaccine
## 3     8 moderate   placebo
## 4    18 moderate   vaccine
## 5     5    large   placebo
## 6    11    large   vaccine
```

```r
dim(vaccine)
```

```
## [1] 6 3
```

```r
glm.null <- glm(count ~ 1 , family = poisson(log), data=vaccine)
glm.null$df.null
```

```
## [1] 5
```

```r
glm.min <- glm(count ~ treatment , family = poisson(log), data=vaccine)
glm.min
```

```
##
## Call:  glm(formula = count ~ treatment, family = poisson(log), data = vaccine)
##
## Coefficients:
##      (Intercept)  treatmentvaccine
##          2.53897          -0.08224
##
## Degrees of Freedom: 5 Total (i.e. Null);  4 Residual
## Null Deviance:       23.81
## Residual Deviance: 23.68     AIC: 52.81
```

```r
glm.min$df.residual
```

```
## [1] 4
```

```r
qchisq(0.95,4)
```

```
## [1] 9.487729
```

- homogeneity model (A+B)

```r
head(vaccine)
```

```
##   count response treatment
## 1    25    small   placebo
## 2     6    small   vaccine
## 3     8 moderate   placebo
## 4    18 moderate   vaccine
## 5     5    large   placebo
## 6    11    large   vaccine
```

```r
glm.homo <- glm(count ~ response + treatment , family = poisson(log), data=vaccine)
glm.homo$df.residual
```

```
## [1] 2
```

```r
glm.homo$deviance
```

```
## [1] 18.64253
```

```r
qchisq(0.95,4)
```

```
## [1] 9.487729
```

- not much of an improvement on 2 degrees of freedom

```
glm.min$deviance - glm.homo$deviance
```

```
## [1] 5.041382
```

```
qchisq(0.95,2)
```

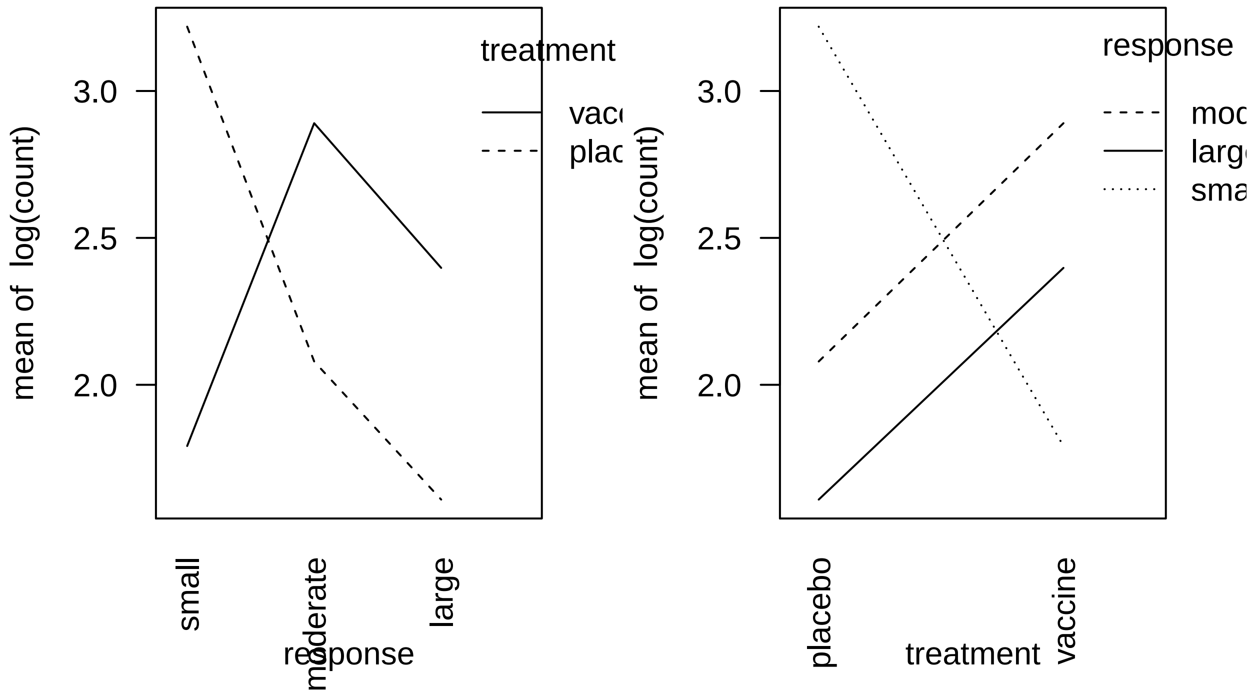
```
## [1] 5.991465
```

- groups differ in their response

Again `interaction.plot` can be used.

```
par(mfrow=c(1,2))
with(vaccine, {
interaction.plot(response, treatment, log(count),las=2)
interaction.plot(treatment, response, log(count),las=2)

})
```



Task 25 verify analysis in Example 6.7

Task 26 What is the largest pearson residual for the A+B model?