troutegg

Heejung Shim

Boxes of trout eggs were buried at five different stream locations and retrieved at 4 different times. The number of surviving eggs was recorded. Question of interest is whether survival of trout eggs depending on time and location.

```r
library(faraway)
data(troutegg)
?troutegg
str(troutegg)
```

```
## 'data.frame':    20 obs. of  4 variables:
##  $ survive : int  89 106 119 104 49 94 91 100 80 11 ...
##  $ total   : int  94 108 123 104 93 98 106 130 97 113 ...
##  $ location: Factor w/ 5 levels "1","2","3","4",..: 1 2 3 4 5 1 2 3 4 5 ...
##  $ period  : Factor w/ 4 levels "4","7","8","11": 1 1 1 1 1 2 2 2 2 2 ...
```

Look at data and check for interaction between location and period.

```r
ftable(xtabs(cbind(survive,total) ~ location + period, troutegg))
```

```
##                  survive total
## location period
## 1        4            89    94
##          7            94    98
##          8            77    86
##          11          141   155
## 2        4           106   108
##          7            91   106
##          8            87    96
##          11          104   122
## 3        4           119   123
##          7           100   130
##          8            88   119
##          11           91   125
## 4        4           104   104
##          7            80    97
##          8            67    99
##          11          111   132
## 5        4            49    93
##          7            11   113
##          8            18    88
##          11            0   138
```
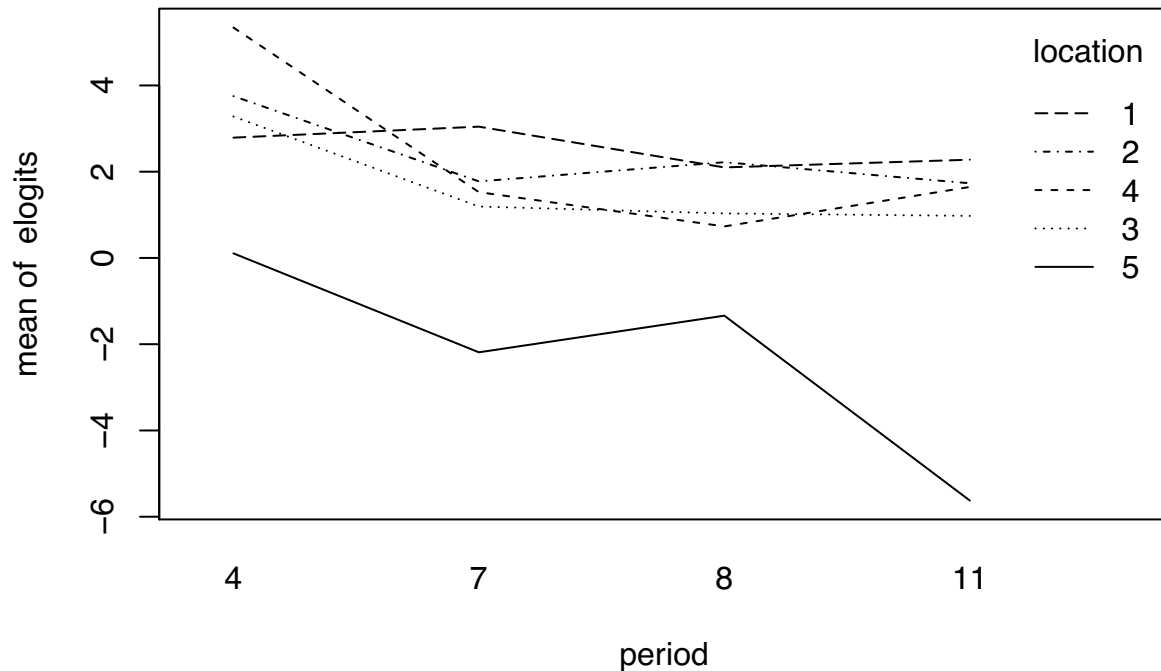
```r
elogits <- log((troutegg$survive+0.5)/(troutegg$total-troutegg$survive+0.5))
with(troutegg, interaction.plot(period, location, elogits))
```

(handwritten annotations in right margin:)

$\Rightarrow$ # of survival in total

$\Rightarrow$ use binomial

$Y_i \sim Bin(m_i, p_i)$

$\quad \hookrightarrow$ location, period.

$g(\mu_i) = x_i \beta$

$\downarrow$

$g(y_i) = \log \dfrac{y_i + 0.5}{m_i - y_i + 0.5} \rangle$ avoid $y_i = 0$

```
# no obvious interaction
```

Fit a binomial regression with logit link. Residual deviance is very high - something is wrong with the model!

```
bmod <- glm(cbind(survive,total-survive) ~ location+period, family=binomial, troutegg)
summary(bmod)
```

```
##
## Call:
## glm(formula = cbind(survive, total - survive) ~ location + period,
##     family = binomial, data = troutegg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8305  -0.3650  -0.0303   0.6191   3.2434
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.6358     0.2813  16.479  < 2e-16 ***
## location2    -0.4168     0.2461  -1.694   0.0903 .
## location3    -1.2421     0.2194  -5.660 1.51e-08 ***
## location4    -0.9509     0.2288  -4.157 3.23e-05 ***
## location5    -4.6138     0.2502 -18.439  < 2e-16 ***
## period7      -2.1702     0.2384  -9.103  < 2e-16 ***
## period8      -2.3256     0.2429  -9.573  < 2e-16 ***
## period11     -2.4500     0.2341 -10.466  < 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: 1021.469  on 19  degrees of freedom
## Residual deviance:   64.495  on 12  degrees of freedom
## AIC: 157.03
```
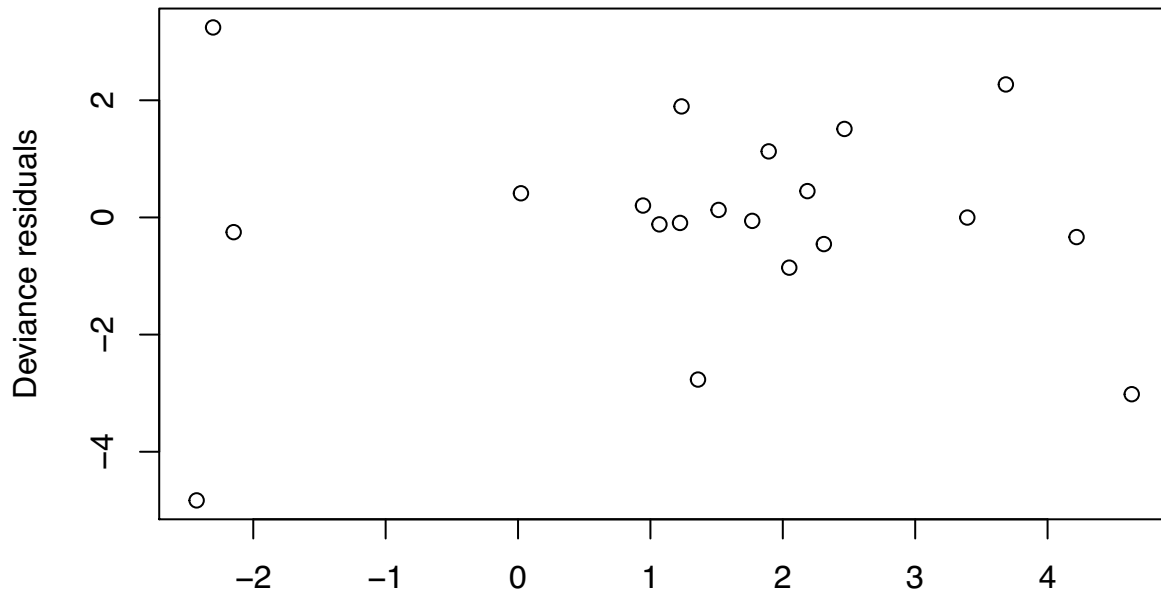
```
## 
## Number of Fisher Scoring iterations: 5
```

(handwritten left margin: check the scaled deviance)

```
pchisq(deviance(bmod),12, lower.tail=F)
```

```
## [1] 3.379416e-09
```

(handwritten right: n=20 → 5 levels, so 4 parameters
p = 1+4+3 = 8
↳ 4 levels periods so 3 parameters.)

Check residuals - no trends.

```
plot(residuals(bmod) ~ predict(bmod,type="link"),
     xlab=expression(hat(eta)), ylab="Deviance residuals")
```


(handwritten: checking overdispersion for binomial and poisson regression
$\hat{\phi} \approx 1$ if no overdisperson)

Let's estimate phi - should be close to 1 for a binomial regression, but isn't!

```
n <- 20
p <- 8
(phihat <- sum(residuals(bmod,type="pearson")^2)/(n-p))
```

```
## [1] 5.330322
```

Fitting quasibinomial model.

```
# fitting quasibinomial model
qbmod <- glm(cbind(survive,total-survive) ~ location+period, family=quasibinomial, troutegg)
summary(qbmod)
```

```
## 
## Call:
## glm(formula = cbind(survive, total - survive) ~ location + period,
##     family = quasibinomial, data = troutegg)
## 
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8305  -0.3650  -0.0303   0.6191   3.2434
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    4.6358     0.6495    7.138 1.18e-05 ***
## location2      -0.4168     0.5682   -0.734 0.477315
## location3      -1.2421     0.5066   -2.452 0.030501 *
## location4      -0.9509     0.5281   -1.800 0.096970 .
## location5      -4.6138     0.5777   -7.987 3.82e-06 ***
## period7        -2.1702     0.5504   -3.943 0.001953 **
## period8        -2.3256     0.5609   -4.146 0.001356 **
## period11       -2.4500     0.5405   -4.533 0.000686 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 5.330358)
##
##      Null deviance: 1021.469  on 19  degrees of freedom
## Residual deviance:   64.495  on 12  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

`0.2461*sqrt(phihat)`

```
## [1] 0.5681831
```

Fitting quasibinomial model by rescaling variances by phihat

`summary(bmod, dispersion=phihat)`

```
##
## Call:
## glm(formula = cbind(survive, total - survive) ~ location + period,
##     family = binomial, data = troutegg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8305  -0.3650  -0.0303   0.6191   3.2434
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.6358     0.6495   7.138 9.49e-13 ***
## location2    -0.4168     0.5682  -0.734   0.4632
## location3    -1.2421     0.5066  -2.452   0.0142 *
## location4    -0.9509     0.5281  -1.800   0.0718 .
## location5    -4.6138     0.5777  -7.987 1.39e-15 ***
## period7      -2.1702     0.5504  -3.943 8.05e-05 ***
## period8      -2.3256     0.5609  -4.146 3.38e-05 ***
## period11     -2.4500     0.5405  -4.533 5.82e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 5.330322)
##
##      Null deviance: 1021.469  on 19  degrees of freedom
## Residual deviance:   64.495  on 12  degrees of freedom
## AIC: 157.03
##
## Number of Fisher Scoring iterations: 5
```

sd ↑

p-value less significant

$$\text{var}(\hat{\beta}) = (x^T \Sigma^{-1} x)^{-1}$$
$$\Sigma_{ii} = g'(\mu_i)^2 V(\mu)$$
for quasibinomial
$$\text{var}(\hat{\beta}^Q) = (x^T \Sigma^{-1} x)^{-1}$$
$$\Sigma_{ii} = g'(\mu_i)^2 V(\mu_i) \phi$$
$$\text{var}(\hat{\beta}^Q) = \text{var}(\hat{\beta}) \phi$$
$$sd(\hat{\beta}^Q) = sd(\hat{\beta}) \sqrt{\hat{\phi}}$$

Test for significance of location.

```r
qbmod2 <- glm(cbind(survive,total-survive) ~ period, family=quasibinomial, troutegg)
summary(qbmod2)
```

```
##
## Call:
## glm(formula = cbind(survive, total - survive) ~ period, family = quasibinomial,
##     data = troutegg)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -17.3778    0.8058    3.0192    4.5949    7.1826
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.139      1.070   1.999   0.0629 .
## period7       -1.333      1.277  -1.044   0.3118
## period8       -1.336      1.298  -1.029   0.3186
## period11      -1.453      1.233  -1.178   0.2561
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 56.33299)
##
##     Null deviance: 1021.47  on 19  degrees of freedom
## Residual deviance:  913.56  on 16  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```r
anova(qbmod, qbmod2, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(survive, total - survive) ~ location + period
## Model 2: cbind(survive, total - survive) ~ period
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1        12      64.50
## 2        16     913.56 -4  -849.06 39.822 7.779e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# check calculation of F statistic for testing location
(913.56 - 64.5)/4/phihat
```

```
## [1] 39.82218
```

Or using

```r
bmod2 <- glm(cbind(survive,total-survive) ~ period, family=binomial, troutegg)
anova(bmod, bmod2, dispersion=phihat, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(survive, total - survive) ~ location + period
## Model 2: cbind(survive, total - survive) ~ period
```

```
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1        12      64.50
## 2        16     913.56 -4  -849.06 39.822 7.779e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```