Lecture 6: Poisson, Quasi-Poisson & Negative Binomial

This example will use the “Horseshoe Crab Mating” data from Agresti (Section 4.3.2). Data was collected on 173 female crabs and their mating patterns. During mating, male crabs that cluster around the female are called *satellites*. We are interested in modeling a female crab’s number of *satellites* as a function of the female crab’s *color*, *spine* condition, *weight* and carapace *width*. We can read-in this data from Agresti’s website.

```
rm(list=ls())
data = read.delim("http://users.stat.ufl.edu/~aa/cda/data/Crabs.dat",sep="")
summary(data)
```

```
##       crab          sat               y              weight     
##  Min.   :  1   Min.   : 0.000   Min.   :0.0000   Min.   :1.200  
##  1st Qu.: 44   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:2.000  
##  Median : 87   Median : 2.000   Median :1.0000   Median :2.350  
##  Mean   : 87   Mean   : 2.919   Mean   :0.6416   Mean   :2.437  
##  3rd Qu.:130   3rd Qu.: 5.000   3rd Qu.:1.0000   3rd Qu.:2.850  
##  Max.   :173   Max.   :15.000   Max.   :1.0000   Max.   :5.200  
##      width          color           spine      
##  Min.   :21.0   Min.   :1.000   Min.   :1.000  
##  1st Qu.:24.9   1st Qu.:2.000   1st Qu.:2.000  
##  Median :26.1   Median :2.000   Median :3.000  
##  Mean   :26.3   Mean   :2.439   Mean   :2.486  
##  3rd Qu.:27.7   3rd Qu.:3.000   3rd Qu.:3.000  
##  Max.   :33.5   Max.   :4.000   Max.   :3.000  
```

1 Poisson GLM

```
model = glm(sat ~ weight + width + factor(color) + factor(spine), data=data, family=poisson())
summary(model)
```

```
## 
## Call:
## glm(formula = sat ~ weight + width + factor(color) + factor(spine), 
##     family = poisson(), data = data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.0290  -1.8630  -0.5988   0.9331   4.9446  
## 
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)   
## (Intercept)     -0.36180    0.96655  -0.374  0.70817   
## weight           0.49647    0.16626   2.986  0.00283 **
## width            0.01675    0.04892   0.342  0.73207   
## factor(color)2 -0.26485    0.16811  -1.575  0.11515   
## factor(color)3 -0.51371    0.19536  -2.629  0.00855 **
## factor(color)4 -0.53086    0.22692  -2.339  0.01931 * 
## factor(spine)2 -0.15037    0.21358  -0.704  0.48139   
## factor(spine)3  0.08728    0.11993   0.728  0.46674   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 549.59  on 165  degrees of freedom
## AIC: 920.88
##
## Number of Fisher Scoring iterations: 6
```

```
qchisq(.95,dim(data)[1]-1)
```

```
## [1] 203.6015
```

```
qchisq(.95,dim(data)[1]-length(model$coefficients))
```

```
## [1] 195.9734
```

```
model$null.deviance - model$deviance
```

```
## [1] 83.20607
```

```
qchisq(.95,length(model$coefficients)-1)
```

```
## [1] 14.06714
```

The Fisher Scoring algorithm is used to find the MLEs and converged in 6 iterations.

The *Residual Deviance* compares the reduced model (the fitted model) to the saturated model. The large value of 549.59 indicates lack of fit for the fitted model, i.e. there is a large difference in the loglikelihoods for the reduced vs. saturated models.

The *Null Deviance* compares the intercept only model to the saturated model. The large value of 632.79 indicates lack of fit for the intercept only, i.e. there is a large difference in the loglikelihoods for the reduced vs. saturated models.

The difference between the *Residual Deviance* and *Null Deviance* compares the intercept only model to the reduced model (the fitted model). The large value of 83.21 indicates lack of fit in the intercept only model, i.e. there is a large difference in the loglikelihoods for the intercept only vs. reduced models.

```
exp(model$coefficients)
```

```
##    (Intercept)         weight          width factor(color)2 factor(color)3
##      0.6964214      1.6429116      1.0168897      0.7673201      0.5982748
## factor(color)4 factor(spine)2 factor(spine)3
##      0.5880989      0.8603880      1.0912050
```

The coefficient estimates provide information about the relationship between *satellites* and the covariates. For example, 1.64 is the multiplicative effect on $\hat{\mu}$ for a one unit increase in *weight*.

2 Quasi-Poisson Model

```
model2 = glm(sat ~ weight + width + factor(color) + factor(spine), data=data, family=quasipoisson())
summary(model2)
```

```
##
## Call:
## glm(formula = sat ~ weight + width + factor(color) + factor(spine),
##     family = quasipoisson(), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0290  -1.8630  -0.5988   0.9331   4.9446
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.36180    1.73852  -0.208   0.8354
## weight          0.49647    0.29905   1.660   0.0988 .
## width           0.01675    0.08799   0.190   0.8493
## factor(color)2 -0.26485    0.30238  -0.876   0.3824
## factor(color)3 -0.51371    0.35139  -1.462   0.1457
## factor(color)4 -0.53086    0.40815  -1.301   0.1952
## factor(spine)2 -0.15037    0.38415  -0.391   0.6960
## factor(spine)3  0.08728    0.21571   0.405   0.6863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 3.235255)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 549.59  on 165  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

AIC is *NA* because this model assumes a quasi-likelihood not a true likelihood.

Two important things to note: (1) estimate of the dispersion parameter, and (2) the standard errors for the coefficients.

The estimated dispersion parameter of 3.24 is larger than 1 indicating overdispersion.

The coefficient estimates (and therefore residuals and deviances) are the same as in the Poisson GLM, but the standard errors of the estimates are much larger when the overdispersion is accounted for.

3 Negative Binomial Model

```
library(MASS)
model3 = glm.nb(sat ~ weight + width + factor(color) + factor(spine), data=data)
summary(model3)
```

```
##
## Call:
## glm.nb(formula = sat ~ weight + width + factor(color) + factor(spine),
##     data = data, init.theta = 0.9649768526, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8789  -1.3685  -0.3264   0.4224   2.2292
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.277348   1.950163  -0.142   0.8869
## weight          0.700891   0.356475   1.966   0.0493 *
## width          -0.002437   0.099663  -0.024   0.9805
## factor(color)2 -0.320756   0.372720  -0.861   0.3895
## factor(color)3 -0.596268   0.417349  -1.429   0.1531
## factor(color)4 -0.579003   0.466470  -1.241   0.2145
## factor(spine)2 -0.242703   0.398367  -0.609   0.5424
## factor(spine)3  0.042779   0.248431   0.172   0.8633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.965) family taken to be 1)
##
##     Null deviance: 220.67  on 172  degrees of freedom
## Residual deviance: 196.51  on 165  degrees of freedom
## AIC: 763.32
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.965
##           Std. Err.:  0.176
##
##  2 x log-likelihood:  -745.320
```

```
qchisq(.95,dim(data)[1]-length(model3$coefficients))
```

```
## [1] 195.9734
```

```
1 - pchisq(model3$deviance,dim(data)[1]-length(model3$coefficients))
```

```
## [1] 0.04733231
```

```
qchisq(.95,dim(data)[1]-1)
```

```
## [1] 203.6015
```

```
model3$null.deviance - model3$deviance
```

```
## [1] 24.15588
```

```
qchisq(.95,length(model3$coefficients)-1)
```

```
## [1] 14.06714
```

The *Residual Deviance* compares the reduced model (the fitted model) to the saturated model. The value of 196.51 indicates borderline evidence of lack of fit: the p-value for rejecting the null of the reduced model is just below .05. The null is not rejected at $\alpha = .01$.

The standard errors are slightly larger than in the quasi-Poisson model.

The estimate of the dispersion parameter $\theta = 0.96$ is consistent with overdispersion ($\gamma = 1/\theta = 1.04$).

Note: the *glm.nb* function iterates between the glm fitting and estimating θ. The negative binomial distribution is only an exponential family distribution when $\gamma = 1/\theta$ is known.