Clustered data models - Exercises 1

(Problems from 1 to 5 have been taken from Agresti, 2015, and problem 6 from Faraway, 2016)

1. (7.2 in Agresti) Suppose y_i are independent Poisson variates, with $\mu = \mathrm{E}(y_i)$, $i = 1, ..., n$. For testing $H_0: \mu = \mu_0$, show that the likelihood ratio statistic simplifies to

$$-2(L_0 - L_1) = 2[n(\mu_0 - \bar{y}) + n\bar{y}\log(\bar{y}/\mu_0)].$$

(Here, $L_0 = \log p(\mathbf{y}; \mu_0)$ and $L_1 = \log p(\mathbf{y}; \mu)$ denote log-likelihood functions.)

2. (5.10 in Agresti) The calibration problem is that of estimating x_0 at which $P(y = 1) = \pi_0$ for some fixed π_0, such as 0.5. For the logistic model with a single explanatory variable, explain why a confidence interval for x_0 is the set of x values for which

$$|\hat{\beta}_0 + \hat{\beta}_1 x - \mathrm{logit}(\pi_0)|/[\mathrm{Var}(\hat{\beta}_0) + x^2\mathrm{Var}(\hat{\beta}_1) + 2x\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)]^{1/2} < z_{\alpha/2}.$$

How could you invert a likelihood-ratio test to form an interval?

3. (7.32 in Agresti) For the horseshoe crab data, the negative binomial modeling shown in the R output first treats color as nominal-scale and then in quantitative manner, with the category numbers as scores. Interpret the result of the likelihood ratio test comparing the two models. For the simpler model, interpret the color effect and interpret results of the likelihood test of the null hypothesis of no color effect.

```
library(MASS)
fit.nb.color <- glm.nb(y ~ factor(color), data = Crabs) # Using Crabs.dat file
summary(fit.nb.color)
##
## Call:
## glm.nb(formula = y ~ factor(color), data = Crabs, init.theta = 0.8018786143,
##     link = log)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.4069     0.3526   3.990 6.61e-05 ***
## factor(color)2  -0.2146     0.3750  -0.572    0.567
## factor(color)3  -0.6061     0.4036  -1.502    0.133
## factor(color)4  -0.6913     0.4508  -1.533    0.125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.8019) family taken to be 1)
##
##     Null deviance: 199.23  on 172  degrees of freedom
## Residual deviance: 194.00  on 169  degrees of freedom
## AIC: 772.3
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.802
##           Std. Err.:  0.136
##
##  2 x log-likelihood:  -762.296
```

```r
fit.nb.color2 <- glm.nb(y ~ color, data = Crabs) # Using color scores 1,2,3,4
summary(fit.nb.color2)
##
## Call:
## glm.nb(formula = y ~ color, data = Crabs, init.theta = 0.798572811,
##     link = log)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.7045     0.3095   5.507 3.66e-08 ***
## color        -0.2689     0.1225  -2.194   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.7986) family taken to be 1)
##
##     Null deviance: 198.77  on 172  degrees of freedom
## Residual deviance: 193.94  on 171  degrees of freedom
## AIC: 768.68
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.799
##           Std. Err.:  0.136
##
##  2 x log-likelihood:  -762.679
```

```r
anova(fit.nb.color2, fit.nb.color)
## Likelihood ratio tests of Negative Binomial Models
##
## Response: y
##           Model     theta Resid. df    2 x log-lik.   Test    df LR stat.
## 1         color 0.7985728       171       -762.6794
## 2 factor(color) 0.8018786       169       -762.2960 1 vs 2     2 0.383383
##     Pr(Chi)
## 1
## 2 0.8255615
```

```r
1 - pchisq(767.409 - 762.679, df = 172 - 171) # LR test vs. null model
## [1] 0.02964089
```

4. (7.33 in Agresti) For the horseshoe crab data, the following output shows a zero-inflated negtive binomial model using quantitative color for the zero component. Interpret results, and compare with the NB2 model fitted in the previous exercise with quantitative color. Can you conduct a likelihood-ratio test comparing them? Why or why not?

```r
library(pscl)
summary(zeroinfl(y ~ 1 | color, dist = "negbin", data = Crabs))
##
## Call:
## zeroinfl(formula = y ~ 1 | color, data = Crabs, dist = "negbin")
##
```

```
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.2246 -0.8186 -0.1712  0.5465  3.7220
##
## Count model coefficients (negbin with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.46324    0.06892  21.231  < 2e-16 ***
## Log(theta)   1.47997    0.35114   4.215  2.5e-05 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.7520     0.6658  -4.133 3.58e-05 ***
## color         0.8023     0.2389   3.358 0.000785 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 4.3928
## Number of iterations in BFGS optimization: 14
## Log-likelihood:  -363 on 4 Df
```

5. (5.32 in Agresti) For the horseshoe crab dataset, let $y = 1$ if a female crab has at least one satellite, and let $y = 0$ if a female crab does not have any satellites. Fit a main-effects logistic model using color and weight as explanatory variables. Interpret and show how to conduct inference about the color and weight effects. Next, allow interaction between color and weight in their effects on y, and test whether this model provides a significantly better fit.

6. (3.2 in Faraway) Incubation temperature can affect the sex of turtles. An experiment was conducted with three independent replicates for each temperature and the number of male and female turtles born was recorded and can be found in the `turtle` dataset, included in the R package `faraway`.

(a) Plot the proportion of males against the temperature. Comment on the nature of the relationship.
(b) Fit a binomial response model with a linear term in temperature. Does this model fit the data?
(c) Check for outliers.
(d) Add a quadratic term in temperature. Is this additional term a significant predictor of the response? Does the quadratic model fit the data?
(e) There are three replicates for each value of temperature. Assuming independent binomial variation, how much variation would be expected in the three proportions observed? Compare this to the observed variation in these proportions. Do they approximately agree or is there evidence of greater variation?
(f) If the three replicates are homogenous, they could be combined so that the dataset would have only five cases in total. Create this dataset and fit a model linear in temperature. Compare the fit seen for this model with that found in (b).