Clustered data models - Exercises 2

1. (8.2 in Agresti) Using $\mathrm{E}(y) = \mathrm{E}[\mathrm{E}(y|x)]$ and $\mathrm{Var}(y) = \mathrm{E}[\mathrm{Var}(y|x)] + \mathrm{Var}[\mathrm{E}(y|x)]$, derive the mean and variance of the beta-binomial distribution.

2. (8.3 in Agresti) Let y_1 and y_2 be independent and identically distributed negative binomial variates with dispersion parameter γ. (See the definition of γ in Section 7.3.2 of Agresti).

   a) Show that $y_1 + y_2$ is negative binomial with dispersion parameter $\gamma/2$.
   b) Conditional on $y_1 + y_2$, show that y_1 has a beta-binomial distribution.
   c) State the multicategory extension of (b) that yields a Dirichlet-multinomial distribution. Explain the analogy with the corresponding Poisson-multinomial result (Section 7.2.1. in Agresti).

3. (8.6 in Agresti) Motivation for the quasi-score equations (Equation 8.2 in Agresti): suppose we replace $\nu(\mu_i)$ by known variance ν_i. Show that the equations result from the weighted least squares approach of minimizing $\sum_i [(y_i - \mu_i)^2/\nu_i]$.

4. (7.31 in Agresti) The table below (Table 7.5 in Agresti; the data available at www.stat.ufl.edu/~aa/glm/data), summarizes responses of 1308 subjects to the question: within the past 12 months, how many people have you known personally that were victims of homicide? The table shows responses by race, for those who identified their race as white or as black.

   a) Let y_i denote the response for subject i and let $x_i = 1$ for blacks and $x_i = 0$ for whites. Fit the Poisson GLM $\log \mu_i = \beta_0 + \beta x_i$ and interpret $\hat{\beta}$.
   b) Describe factors of heterogeneity such that a Poisson GLM may be inadequate. Fit the corresponding negative binomial GLM, and estimate how the variance depends on the mean. What evidence does this model fit provide that the Poisson GLM had overdispersion? (Table 7.5 also shows the fits for these two models.)
   c) Show that the Wald 95% confidence interval for the ratio of means for blacks and whites is (4.2, 7.5) for the Poisson GLM but (3.5, 9.0) for the negative binomial GLM. Which do you think is more reliable? Why?

```r
homic <- read.table("../data/Homicides.dat", header = TRUE)

# Frequency table
n <- table(homic$race)
tb <- table(homic$count, homic$race)
colnames(tb) <- c("white", "black")

# Fitted values for the Poisson distribution
mod.pois <- glm(count ~ race, family = poisson, homic)
mu.pois.black <- exp(sum(mod.pois$coef))
mu.pois.white <- exp(mod.pois$coef[1])
tb.pois <- round(cbind(n[1]* dpois(0:6, mu.pois.white),
                       n[2] * dpois(0:6, mu.pois.black)), 1)

# Fitted values for the negative binomial distributin
library(MASS)
mod.nb <- glm.nb(count ~ race, homic)
mu.nb.black <- exp(sum(mod.nb$coef))
mu.nb.white <- exp(mod.nb$coef[1])
```

```r
tb.nb <- round(cbind(n[1]* dnbinom(0:6, mu = mu.nb.white, size = mod.nb$theta),
                     n[2] * dnbinom(0:6, mu = mu.nb.black, size = mod.nb$theta)), 1)
dimnames(tb.nb) <- dimnames(tb.pois) <- dimnames(tb)

# Compare Table 7.5 in Agresti
cbind(tb, tb.pois, tb.nb)
##   white black  white black  white black
## 0  1070   119 1047.7  94.3 1064.9 122.8
## 1    60    16   96.7  49.2   67.5  17.9
## 2    14    12    4.5  12.9   12.7   7.8
## 3     4     7    0.1   2.2    2.9   4.1
## 4     0     3    0.0   0.3    0.7   2.4
## 5     0     2    0.0   0.0    0.2   1.4
## 6     1     0    0.0   0.0    0.1   0.9
```

5. (8.14 in Agresti) Use QL methods to analyze Table 7.5 on counts of homicide victims. Interpret, and compare results with Poisson and negative binomial GLMs.

6. (8.13 in Agresti) Use QL methods to construct a model for the horseshoe crab satellite counts, using weight, color, and spine condition as explanatory variables. Compare results with those obtained with zero-inflated GLMs (Section 7.5 in Agresti).

7. (8.16 in Agresti) For the teratology study analyzed in Section 8.2.4, analyze the data using only the group indicators as explanatory variables (i.e., ignoring hemoglobin). Interpret results. Is it sufficient to use the simpler model having only the placebo indicator for the explanatory variable?