Exercise #3
Due: January 27, 2010

Bio/Stat 571
Winter 2010
P. Heagerty

Reading:
- Cameron & Trivedi, Chapter 2 (skip 2.5.2-2.5.4)
- Liang & McCullagh (1993) *Biometrics* article
- McCullagh & Nelder, sections 4.5, 6.2 (overdispersion)
- Lindsey, Chapter 5, "Overdispersion"
- Carroll, Ruppert & Stefanski, Appendix A

1. The file `Bees-data.txt` contains information about the "working activity" of bees as a function of time of day. One of the important characteristics of working activities is the number of bees leaving the hive for outside activity. The data record the number of bees that left the hive per hour, and the hour under study. The data contian the bee count (`Number`) over 11 hours in the day (`Time`), and over several successive non-rainy days.

(a) Use poisson regression with different functions of `Time` and evaluate their fit to the data using plots of the fitted mean, and appropriate residual plots. In particular evaluate the following mean models

model 1	linear in `Time`
model 2	quadratic in `Time`
model 3	cubic in `Time`
model 4	linear splines in `Time` with knots at 9, 11, 13, 15
model 5	cubic splines in `Time` with knots at 9, 11, 13, 15
model 6	natural splines in `Time` with knots at 9, 11, 13, 15
model 7	saturated model using `factor(Time)`

Based on these plots which model(s) appear appropriate in order to characterize the average number of bees leaving the hive as a function of hour of day?

(b) Using an appropriate mean model based on your work in 1(a) create residual plots that can summarize evidence for overdispersion. In particular, do these data suggest either a scale form, or a negative binomial form for overdispersion?

(c) Using a negative binomial model fit models 1-7 above and compute the AIC value. Based on these summaries suggest a plausible parsimonious model and plot the fitted mean as a function of time with pointwise 95% confidence intervals.

(d) Compare the fitted mean function (and confidence interval) obtained using appropriate quasi-likelihood methods to the results obtained in 1(c).

(*optional*) Using appropriate methods construct pointwise 90% prediction intervals for the number of bees leaving the hive as a function of time. State the assumptions that you are making in order to obtain these prediction intervals.

2. In this exercise we will consider a mechanism for overdispersed Poisson data and evaluate the resulting impact on standard errors for regression estimators. In particular, we will consider a model for count data when there may be a large number of zero counts. Such models are particularly useful for studies of health care utilization where many subjects have no in-patient (or out-patient) visits while other subjects have 1 or more visits (during a fixed follow-up period). For an agricultural application Hall (2000) *Biometrics* describes a scenario where the impact of chemical treatment is assessed for plants. In some leaves insect reproduction is completely suppressed, while for other leaves the number of insect offspring may be reduced but not completely suppressed. Resulting leaf counts of insect offspring would contain a mixture of zero counts and possibly Poisson counts.

Consider the following hierarchical, or mixture model:

$$\begin{aligned} \nu_i &\sim \text{binomial}(1, p_i) \\ Y_i \mid \nu_i &\sim \text{Poisson}(\ \lambda_i \nu_i\ ) \end{aligned}$$

Note: page 32 of Cameron & Trivedi is particularly helpful for the questions that follow.

(a) Derive the mean and variance of Y_i and express the variance as a function of the mean $\mu_i = E(Y_i)$ and the parameter p_i.

(b) Suppose that we are interested in the mean of Y_i as a function of covariates $\boldsymbol{X}_i$. Assume that both λ_i and p_i may depend on the value of covariates. Since the outcome is a count, it would be possible (and natural) to use Poisson regression methods to link covariates to the mean response. For Poisson regression with log link the score equations are $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{0}$. Using the results of White (1982), what is the variance of the regression estimator, $\widehat{\boldsymbol{\beta}}$, that solves these equations if the variance of Y_i is actually given by your expression in (a)? Note: the matrix A is simply the information matrix assuming the model is correctly specified (ie. if $\text{var}(Y_i) = \mu_i$: the Poisson variance function).

(c) What is the variance of the estimator that solves $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}$ if the true variance of Y_i is given as $\text{var}(Y_i) = \phi\mu_i$?

(d) What is the variance of the estimator that solves $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}$ if the true variance of Y_i is given as $\text{var}(Y_i) = \mu_i + \alpha\mu_i^2$, one parameterization of the negative-binomial variance?

(e) What assumptions regarding μ_i and p_i as functions of $\boldsymbol{X}_i$ would lead to (c), and similarly what assumptions would lead to (d)?

(f) Simulate overdispersed data where $X_{i,1} = i/n$, $X_{i,2} = 0$ if $i <= n/2$ and $X_{i,2} = 1$ if $i > n/2$, and:

$$\begin{aligned} \log \lambda_i &= \gamma_0 + \gamma_1 X_{i,1} + \gamma_2 X_{i,2} \\ \log p_i &= \delta_0 + \delta_1 X_{i,1} \end{aligned}$$

where $\gamma_0 = 0, \gamma_1 = 2, \gamma_2 = 1$, and $\delta_0 = \log(0.60), \delta_1 = \log(0.80/0.60)$. Use a sample size of $N = 150$. Construct residual plot(s) that can be used to guide whether the negative-binomial variance form, $\text{var}(Y_i) = \mu_i + \alpha \cdot \mu_i^2$, or the scale overdispersion model, $\text{var}(Y_i) = \phi \cdot \mu_i$, is suggested by the residuals. Create a plot that can be used to determine if $\text{var}(Y_i) = \mu_i + \alpha_i \cdot \mu_i^2$ where α_i is a function of X_i.

(g) Program your variance estimates given in (c) and (d). For your simulated data set compare the resulting standard error estimates to those obtained from the Poisson assumption. Note: you will need estimates of ϕ and α – use simple moment estimators based on the Pearson residuals.

(h) Program the empirical standard error estimator for this count regression model (ie. log link Poisson regression). For your simulated data set compare the empirical standard error estimates to those obtained in (g).

(i) Simulate 1000 data sets and obtain $\widehat{\boldsymbol{\beta}}$ and the (4) common standard error estimates (ie. assuming Poisson, assuming $\text{var}(Y_i) = \phi\mu_i$, assuming $\text{var}(Y_i) = \mu_i + \alpha\mu_i^2$, and empirical). Compare the standard errors estimates, evaluating whether they appear to be approximately unbiased, and whether they yield nominal coverage for 95% confidence intervals.

(j) **Also**, on your 1000 simulated data sets, use a negative binomial model to estimate $\boldsymbol{\beta}$ and the standard error of $\widehat{\boldsymbol{\beta}}$ (ie. use `glm.nb`). Is the model based standard error provided using the negative binomial model approximately unbiased, and does it yield appropriate coverage? Is there evidence that the negative binomial estimator is more efficient than the poisson regression estimator?

(*optional*) Propose a method for separately estimating the effect of covariates on p_i and λ_i.