Homework 2, Due September 27 at 11:45am

Provide clearly documented code in an R markdown document where appropriate. Whenever possible, provide statistical evidence to support your answers.

1. Agresti 4.5 (30 points)

Consider the following artificial data. When x=0, we have 1 success out of 4 trials. When x=1, we have 2 successes out of 4 trials; when x=2, we have 4 successes out of 4 trials.

Use M_0 to denote an intercept only logistic regression model, and use M_1 to denote the model with an intercept and a linear term in x. Denote the maximized log-likelihood values by L_0 for M_0, L_1 for M_1, and L_s for the saturated model. Consider the data in two formats: (i) ungrouped data, with n_i=1 for $i = 1, 2, ..., 12$, and (ii) grouped data with $n_i = 4$ for $i=1,2,3.

1a.

Fit M_0 and M_1 for each data format, and report L_0 and L_1 in each case. Note the data format does not affect the values of L_0 and L_1.

1b.

Show that the deviances for M_0 and M_1 differ for the two data formats. Why is this? (Hint: Consider the number of parameters in the saturated model for each case.)

1c.

Show that the difference between the deviances for M_0 and M_1 is the same regardless of data format. Why is this? This result implies that for testing for the effect of x, the data format does not matter, though it does matter for testing goodness of fit versus the saturated model.

2. Cellular differentiation (20 points)

Using the cellular differentiation data from class, fit a Poisson regression model with IFN as a nominal categorical variable and interaction terms between IFN and TNF (a saturated model). Assess whether the interaction terms provide significant improvement to the model and diagnose any problems encountered in estimation. Using the main effects model, evaluate whether IFN is best treated as a nominal categorical variable or by fitting a linear trend in IFN dose.

3. Negative binomial distribution (20 points)

3a. Negative binomial as Poisson-Gamma mixture

Consider the negative binomial distribution,

$$f(y \mid p, r) = \binom{r+y-1}{y} p^r (1-p)^y.$$

Suppose we have a Poisson random variable Y where $y \mid \lambda \sim$ Poisson and $\lambda \sim$ Gamma(α, β). Show the negative binomial distribution can be derived as a Poisson-gamma mixture by integrating λ out of the joint likelihood $Pr(y, \lambda) = Pr(y \mid \lambda)Pr(\lambda)$.

Solution

$$
\begin{aligned}
f(y \mid p, r) &= \binom{r+y-1}{y} p^r (1-p)^y \\
p(y \mid \lambda) = \frac{e^{-\lambda}\lambda^y}{y!} \quad \pi(\lambda) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp -\beta\lambda \\
Pr(y, \lambda) &= \frac{e^{-\lambda}\lambda^y}{y!} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp\{-\beta\lambda\} \\
Pr(y \mid \alpha, \beta) &= \int_\lambda \frac{e^{-\lambda}\lambda^y}{y!} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp\{-\beta\lambda\} d\lambda \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)y!} \int_\lambda \lambda^{y+\alpha-1} \exp\{-\lambda(\beta+1)\} d\lambda \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)y!} \frac{\Gamma(y+\alpha)}{(\beta+1)^{y+\alpha}} \int_\lambda \frac{(\beta+1)^{y+\alpha}}{\Gamma(y+\alpha)} \lambda^{y+\alpha-1} \exp\{-\lambda(\beta+1)\} d\lambda \\
&= \frac{\beta^\alpha}{(\beta+1)^y(\beta+1)^\alpha} \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)y!} \\
&= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)y!} (\frac{\beta}{\beta+1})^\alpha (\frac{1}{\beta+1})^y \\
\text{Note that } \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)y!} &= \frac{(y+\alpha-1)(y+\alpha-2)\dots(y+\alpha-y)\Gamma(\alpha)}{\Gamma(\alpha)y!} = \binom{\alpha+y-1}{y} \\
&= \binom{\alpha+y-1}{y} (\frac{\beta}{\beta+1})^\alpha (1-\frac{\beta}{\beta+1})^y
\end{aligned}
$$

So clearly $\alpha = r$ and $p = \dfrac{\beta}{\beta+1}$

3b. Exponential family formulation when r known

Show the negative binomial distribution is a member of the exponential family when r is known but not when r is unknown.

Solution

To begin, I show that when r is known, the negative binomial is in the eponential family. Note that the form of the exponential family is as follows:

$$h(x) \exp\{\eta(\theta)T(X) - A(\eta)\}$$

$$
\begin{aligned}
f(y \mid p, r) &= \binom{r+y-1}{y} p^r (1-p)^y \\
&= \binom{r+y-1}{y} \exp\{r \log p + y \log(1-p)\} \\
\text{Let } \theta = log(1-p) \rightarrow p &= 1 - e^\theta \\
&= \binom{r+y-1}{y} \exp\{r \log(1-e^\theta) + y\theta\} \\
&= \binom{r+y-1}{y} \exp\{y\theta \ - (-r \log(1-e^\theta))\}
\end{aligned}
$$

So note $\eta(\theta) = \log(1-p)$, $T(X) = y$, $A(\eta) = -r\log(1-e^\theta)$, and $h(x) = \binom{r+y-1}{y}$. So when r is known, this is an exponential family. However, when r is not known, this cannot be shown, as follows.

$$
\begin{aligned}
f(y \mid p, r) &= \binom{r+y-1}{y} p^r (1-p)^y \\
&= \exp\{\log(\binom{r+y-1}{y} p^r (1-p)^y)\} \\
&= \exp\{\log \binom{r+y-1}{y} + r \log(p) + y \log(1-p)\} \\
&= \exp\{\log(\frac{(r+y-1)!}{(r-1)!y!})\} \exp\{r \log(p) + y \log(1-p)\} \\
&= \frac{(y+r-1)\ldots(y+1)}{(r-1)!} \exp\{r \log(p) + y \log(1-p)\}
\end{aligned}
$$

4. Hospital Performance (30 points)

The Centers for Medicare and Medicaid Services (CMS) is required by Congress to evaluate hospital performance. We consider data on outcomes of cardiac surgical procedures in New York State from 2008-2011. The outcome of interest is mortality within 30 days of the procedure, and variables available in the dataset include hospital name, procedure type (the CABG procedure is simpler than valve replacement), number of cases/procedures, number of deaths, and expected mortality rate per 100 cases (this variable adjusts for case mix or perceived difficulty of the procedures related to patient characteristics, e.g. some hospitals may see rich, healthy patients while others see patients who have limited resources to devote to medical care). Assume each line in the dataset represents data contributed by a different physician (physician names are not included). The data are in the file cardiacsurgery.csv.

4a. Model fitting and interpretation

Treating each row of the excel spreadsheet as an observation from a binomial distribution with number of failures equal to the number of deaths and number of trials equal to the number of cases/procedures, fit a GLM exploring the contributions of the hospital, region, procedure, and case mix to the mortality rate. Use analysis of deviance to select the most appropriate model for the data, specify your preferred model clearly in equation form, and provide clear interpretations of all parameter estimates.

4b. Hospital rankings

If you are tasked with recognizing "high-achieving" and "low-achieving" hospitals for CMS, which hospitals would you select, and why? Provide a detailed statistical justification for your choice.