Correlated Data Analysis: Modeling, Analytics and Applications

PETER X.-K. SONG

Problem Set 2

Problem 2.1 Consider a gamma distribution with the density function

$$p(y;\psi,\lambda)=\frac{\psi^{\lambda}}{\Gamma(\lambda)}y^{\lambda-1}e^{-\psi y}, y>0.$$

(a) Show the gamma distribution is a reproductive exponential dispersion model.
(b) Give the unit deviance d and the normalizing term a.
(c) Show the unit variance function $V(\mu)=\mu^2$.
(d) Rewrite the gamma distribution in the form of an additive exponential model.
(e) Show that the standard convolution formula for the χ^2-distribution is a special case of the convolution formula for additive ED* model.
(Hint: Find the form of additive model for gamma distribution.)

Problem 2.2 Consider a location-dispersion model

$$p(y;\mu,\sigma^2)=a(\sigma^2)\exp\left\{-\frac{1}{2\sigma^2}d(y-\mu)\right\}, \mu\in\mathcal{R}, \sigma^2>0$$

where $d:\mathcal{R}\to[0,\infty)$ satisfying $d(0)=0$ only.
(a) Is this model a dispersion model? Why?
(b) Consider a special case of $d(y)=|y|^{\rho}$. Show that

$$a(\sigma^2)=\frac{\rho(2\sigma^2)^{-1/\rho}}{2\Gamma(1/\rho)}$$

where $\rho>0$ is given. (Note that $\rho=2$ corresponds to the normal distribution.)
(c) Show that the location-dispersion model defined by

$$d(y)=\log(1+y^2)$$

corresponds to a distribution of $Y=X/\sqrt{\sigma^{-2}-1}+\mu$ where X follows a student's t-distribution with $\sigma^{-2}-1$ degrees of freedom.

Problem 2.3 Let $X_1,\ldots,X_n$ be *i.i.d.* random variables from the *logarithmic distribution* defined by

$$p(y;\rho)=\frac{\rho^y}{-y\log(1-\rho)},\ y=1,2,\ldots.$$

Define

$$S_n = X_1 + X_2 + \ldots + X_n - n.$$

The distribution of $X_1 + \ldots + X_n$ is known as the *Starling distribution.*
(a) Show that S_n follows an additive ED model with dispersion parameter $1/n$.
(b) Give conditions under which $S_n \xrightarrow{d}$ Poisson distribution as $n \to \infty$.

Problem 2.4 Consider a teratological experiment that aims to test if a certain substance is a carcinogen (or toxin). 32 pregnant rats were randomized to different dose levels, with one half being treated by the active substance and the other half treated by a control substance. The number of fetuses born from a common litter survived for 21 day lactation period was reported below:

Treatment group (TG)				Control group (TG)			
12/12	11/11	10/10	9/9	13/13	12/12	9/9	9/9
10/11	9/10	9/10	8/9	8/8	8/8	12/13	11/12
8/9	4/5	7/9	4/7	9/10	9/10	8/9	11/13
5/10	3/6	3/10	0/7	4/5	5/7	7/10	7/10

This is a typical example of aggregation of dependent variables. Let Y_{ij} be the indicator of survival for baby rat j born at litter i, $Y_{ij} = 1$ for survival and 0 for death. Then, the number of survivors Y_i out of n_i at litter i takes the form of sum (or aggregation):

$$Y_i = Y_{i1} + \cdots + Y_{ij} + \cdots + Y_{in_i},\ Y_{ij} = 1 \text{ or } 0,$$

and the resulting Y_i does not follow a binomial distribution, unless $Y_{ij}, j = 1, \ldots, n_i$ are independent and identically distributed. The *litter effect* refers to the tendency for fetuses from the same litter to respond more alike than fetuses from different litters, because of genetic similarity. In this case, the data tend to be overdispersed.

To account for variation in the proportion Y/n among "comparable" litters, assume a beta-binomial distribution defined as follows: Given π_i, $Y_{ij}, j = 1, \ldots, n_i$ are conditionally independent, and

- $Y_{ij}|\pi_i \overset{ind.}{\sim} \text{Binomial}(1, \pi_i), j = 1, \ldots, n_i,$
- $\pi_i \overset{i.i.d.}{\sim} \text{Beta}(\alpha, \gamma), \alpha, \gamma > 0$.

Show the following results:
(a) $\text{E}(Y_i/n_i) = \text{E}(\pi_i)$. Denoted this mean by θ.
(b) $\text{Var}(Y_i/n_i) = \frac{\theta(1-\theta)}{n_i}\{1 + (n_i - 1)\phi\}$, where $\phi = \frac{1}{\alpha+\gamma+1}$. Clearly, the variance is inflated.
(c) $\phi = \text{corr}(Y_{ij}, Y_{ik}), \forall j \neq k$, the common pairwise (Pearson) correlation coefficient between two fetuses from the same litter i.

Problem 2.5 Install R package `CircStats` and use functions in this package to implement the MLE estimation given in Section 2.6.4. Then analyze the following data.

Bats hunt insects by sending out high frequency sounds and listening for echoes. The data below comprises distances and angles observed in a study of the echolocation of flying insects by bats.

Distance (x in cm)	62	52	68	23	34	45	27	42	83	56	40
Angle (y^o)	-15	5	60	50	-15	45	40	-15	35	35	70

(a) Use function `circ.plot` in R to plot the angular data.
(b) Fit the data by the GLM and verify if the covariate of distance, x, is statistically significant to the response of hunting angle, y.
(c) Give the estimates of offset mean parameter μ_0 and dispersion parameter σ^2. Note that function `circ.kappa` in R provides a bias correction for the MLE estimation for the index parameter $\lambda = 1/\sigma^2$.