7. Overdispersion and Correlated Data

- For independent count (Poisoon or Binomial) data, there is often overdispersion. One possible cause is heterogeneity, for example, due to an unobserved important covariate.

- When the data is clustered, ignoring that the correlation can also result in overdispersion which, when ignored, leads to underestimate of the standard errors of the regression parameters.

- Example
  - In teratology experiments, pregnant rats are randomized to receive a teratogenic or a control agent. Then the total number of animals in a litter and the number of birth defects are recorded. Because in a litter all births have the same mother, their outcomes (having birth defect or not) are correlated.
  - Hospitals are randomized to use a new treatment program for alcoholics or stay with the current

program. The response variable in the number of hospitalization in the year following enrollment in the program. One might expect a "hospital" effect such that the hospitalization events at the same hospital are corrected.

- Previously we used quasi-likelihood methods to take into account of the overdispersion by using a scale parameter ϕ.

- For likelihood-based methods, it is natural to model the "litter" or "hospital" effects as random effects. Early example are beta-binomial and Poisson-Gamma (negative binomial) models.

Beta-Binomial Model

- For the teratology experiment, let $Y_{ijk} = 1$ if animal k from litter j in treatment group i has a birth defect, and 0 otherwise. Then we can assume

$$Y_{ijk}|\pi_{ij} \sim^{indep} Bin(1, \pi_{ij}),$$
$$\pi_{ij} \sim^{indep} Beta(\alpha_i, \beta_i).$$

Then Y_{ijk} has a marginal Bernoulli distribution with mean

$$\mu_k = E(Y_{ijk}) = \frac{\alpha_i}{\alpha_i + \beta_i}.$$

Note that $Y_{ij+} = \sum_k Y_{ijk}$ does not have a Binomial distribution, but a Beta-Binomial distribution with density

$$p(Y_{ij+} = y; \alpha_i, \beta_i) = \binom{n_{ij}}{y} \frac{B(y + \alpha_i, n_{ij-y+\beta_i})}{B(\alpha_i, \beta_i)},$$

where B is the beta function. The mean and variance are

$$\begin{aligned} E(Y_{ij+}) &= n_{ij}\mu_i, \\ var(Y_{ij+}) &= n_{ij}\mu_i(1-\mu_i)\left\{1 + \rho_i(n_{ij}-1)\right\}, \end{aligned}$$

where $\rho_i = \frac{1}{\alpha_i+\beta_i+1} = \frac{1}{\theta_i+1}$ and $\mu_i = \frac{\alpha_i}{\alpha_i+\beta_i}$.
Note that the Beta-Binomial model is not in the exponential family. Even though it is adequate for simple experiments but is more difficult to extend, for example, to model covariates or to allow more complicated correlation structure.

- In contrast, in a quasi-likelihood model we may use

$$var(Y_{ij+}) = \phi n_{ij}\mu_i(1-\mu_i) \tag{1}$$

or use a variance function motivated by Beta-Binomial distribution with fixed ρ

$$var(Y_{ij+}) = n_{ij}\mu_i(1-\mu_i)\left\{1+(n_{ij}-1)\rho\right\}. \tag{2}$$

There is no clear conclusion which model is better (Liang and McCullagh, 1993).

Example: Moore's Teratology Data

This data example is from the website of Agresti (2002). Female rats were put on iron-deficient diets and then randomized to receive placebo (group 1) and three different regents of iron supplements (groups 2, 3, and 4). They are sacrificed 3 weeks after pregnant.

n is the total number of fetuses in a litter and y is the number of dead fetuses.

What Sandwich Would You Like?

When the model is misspecified, the asymptotic variance of the MLE has the sandwich form:

$$var(\hat{\beta}) \approx A^{-1}BA^{-1},$$

where

$$A = X^T V^{-1} X,$$
$$B = X^T V^{-1} cov(Y) V^{-1} X.$$

V is the estimated covariance matrix of Y under the current model (and A^{-1} is the "model-based" variance estimate). $cov(Y)$ is the estimated covariance of Y under the "true" model. When the true model is not known, the empirical variance is used.

Empirical variance estimator is a sandwich estimator but not the only one.

Example: Teratology Data

Poisson-Gamma Model

For clustered count data, we can assume

$$Y_{ij}|\mu_{ij} \sim^{ind.} Poisson(\mu_{ij}),$$
$$\mu_{ij} \sim^{ind.} Gamma(\lambda_i, \theta_i/\lambda_i).$$

The marginal distribution of Y_{ij} is negative-binomial with

$$E(Y_{ij}) = E(E(Y_{ij}|\mu_{ij})) = \theta_i,$$
$$var(Y_{ij}) = \theta_i + \theta_i^2/\lambda_i.$$

Note that we use a somewhat unusual parameterization for Gamma distribution such that $E(\mu_{ij}) = \theta_i$ and $var(\mu_{ij}) = \theta_i^2/\lambda_i$. We can then model the marginal mean response as

$$\log(E(Y_{ij})) = x_{ij}^T\beta.$$

For fixed λ_i, the negative binomial is in the exponential family. Therefore, to find the MLE for the negative binomial model, we can alternate the two iterative steps until convergence:

1. For fixed λ, fit the GLM to solve for β in a regression model.

2. For fixed θ_i, estimate λ using Newton-Raphson.

The glm.nb function in R MASS library implements this iterative procedure while gnlr in repeated can estimate all parameters simultaneously.

Conjugated Mixture Model

- The Beta-Binomial and Poisson-Gamma models are examples of conjugated mixture models where the marginal distribution has closed form.

- These models are mainly used to account for overdispersion or simple clustered data and are less suitable to study longitudinal data where the correlation structure may be more complicated.

- When used to model covariates, the computational simplicity of these conjugated mixture models is lost.

- In Lee and Nelder (1991), they extended this class of model to what they called hierarchical generalized linear model which are like generalized linear mixed models but the distribution of the random effects need not be normal. The further extended restricted likelihood estimation to hierarchical or h-likelihood that allows the estimation of the fixed and random effects.