STAT347: Generalized Linear Models
Lecture 11

Today's topics: Chapters 8

- Negative Binomial GLM and Beta-Binomial GLM
- Quasi-likelihood
- Estimating equations and the Sandwich estimator

1 The over-dispersion phenomenon

When we apply the standard GLM models assuming the data are Binomial or Poisson distributed to real data, we always find over-dispersion. Let $v^\star(y_i)$ be the variance of y_i under our model assumption.

- $v^\star(y_i) = n_i p_i(1-p_i)$ for Binomial data and $v^\star(y_i) = \mu_i$ for Poisson counts.
- Over-dispersion: the actual $\mathrm{Var}(y_i) > v^\star(y_i)$.
- We can check whether there is over-dispersion by plotting $\widehat{v^\star}(y_i)$ V.S. $(y_i - \hat{\mu}_i)^2$
- When this happens, it means that our assumption for the randomness of y_i is problematic.

1.1 Negative Binomial distribution for dispersed counts

This is what we have covered in Lecture 10.

- Negative binomial distribution: $y_i \sim \mathrm{Poisson}(\lambda_i)$ and $\lambda \sim \mathrm{Gamma}(\mu_i, k_i)$. Then $y_i \sim \mathrm{NB}(\mu_i, k_i)$
- We have $E(y_i) = \mu_i$ and $\mathrm{Var}(y_i) = \mu_i + \gamma_i \mu_i^2$ where $\gamma_i = 1/k_i$ is the dispersion parameter.
- NB GLM: we assume that $\log(\mu_i) = X_i^T \beta$ and $\gamma_i \equiv \gamma$.

1.2 Beta-Binomial distribution for dispersed Binary data

For the ungrouped Binary data, previous Binary GLM assumed that conditional on having the same X_i, the y_i are i.i.d. Bernoulli trials. But what if the samples are clustered? (Read Chapter 8.2.1).

We may still assume independent grouped data samples, but the individual within each group are allowed to be correlated.

Consider the grouped data. Analogous to the Poisson case, we can have the scenario $y_i \sim \mathrm{Binomial}(n_i, p_i)$ but $\mathrm{logit}(p_i) = X_i^T \beta + \epsilon_i$. We will then have

$$\mathrm{Var}(y_i) > n_i p_i (1 - p_i)$$

The Beta-binomial distribution assumes that $y \sim \mathrm{Binomial}(n, p)$ and $p \sim \mathrm{beta}(\alpha_1, \alpha_2)$. The beta distribution of p has the density function:

$$f(p; \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$$

and

$$E(p) = \mu = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

The Beta-binomial distribution then has the property that

$$E(y) = n\mu, \quad \mathrm{Var}(y) = n\mu(1-\mu)\left[1 + (n-1)\frac{\theta}{1+\theta}\right]$$

where $\theta = 1/(\alpha_1 + \alpha_2)$.

Beta-binomial GLM: we assume the grouped data follows $y_i \sim \text{Beta-binomial}(\mu_i, \theta)$. The relation between μ_i and X_i are the same as we assumed for the standard binary GLM. For example:

$$\mathrm{logit}(\mu_i) = X_i^T \beta$$

Both β and θ are unknown but we can estimate using MLE.

2 Quasi-likelihood

Remind the the score equation for the exponential family distributed data is:

$$\frac{\partial L}{\partial \beta_j} = \sum_i \frac{(y_i - \mu_i)x_{ij}}{\mathrm{Var}(y_i)} \frac{1}{g'(\mu_i)} = 0$$

- These score equations only involve $E(y_i) = \mu_i$ and $\mathrm{Var}(y_i)$.
- We may be OK with the link function, so we are OK with the forms of μ_i and $g'(\mu_i)$ as functions of β.
- However, the form of $\mathrm{Var}(y_i)$ does not fit the data as we see phenomenons like over-dispersion.
- Quasi-likelihood: we replace $\mathrm{Var}(y_i)$ by some other mean-variance relationship that typically involves another unknown dispersion parameter.
- Here, we DO NOT assume any other aspects of the distribution of y_i besides mean and variance.

Common forms of mean-variance relationship $\mathrm{Var}(y_i) = a(\mu_i, \phi)$:

- Proportional: $a(\mu_i, \phi) = \phi v^\star(\mu_i)$.
  - counts: assume $a(\mu_i, \phi) = \phi\mu_i$
  - grouped Binary data: $a(\mu_i, \phi) = \phi\mu_i(n_i - \mu_i)/n_i$
- For counts we can also assume $a(\mu_i, \phi) = \mu_i + \phi\mu_i^2$
- For grouped Binary data we can also assume $a(\mu_i, \phi) = \mu_i(n_i - \mu_i)\left(1 + (n_i - 1)\phi\right)$

Some related properties:

- The proportional mean-variance relationship is the easiest for the computation of $\hat{\beta}$ as ϕ cancels and does not affect solving the "score" equations.
- $\mathrm{Var}(\hat{\beta})$ is affected by ϕ for any of the above mean-variance relationships.
- The purpose to include ϕ is to get the correct estimate of the accuracy of $\hat{\beta}$.

How to estimate ϕ?

- When $a(\mu_i, \phi) = \phi v^\star(\mu_i)$, we can get $\hat{\beta}$ thus $\hat{\mu}_i$ first without knowing ϕ. Then define

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i^2)^2}{\phi v^\star(\hat{\mu}_i)}$$

we can solve ϕ by solving $X^2 = n - p$, which is

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i^2)^2}{v^\star(\hat{\mu}_i)}$$

- For other forms of $a(\mu, \phi)$, we need to solve ϕ and β simultaneously. How to do that?

3 Estimation equations and Sandwich estimator

In quasi-likelihood, we have p equations for β:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{a(\mu_i, \phi)} \frac{1}{g'(\mu_i)} = 0$$

These are not the likelihood score equations, but they are called estimation equations as we solve them to get an estimate of β.

As ϕ is also unknown, we have $p+1$ parameters and we need one more equation for ϕ. Typically, we use the equation

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i^2)^2}{a(\mu_i, \phi)} - (n-p) = 0$$

So now we have $p+1$ equations for $p+1$ parameters.

How to estimate $\mathrm{Var}(\beta)$ without assuming the full distribution of the data?

Consider the general form of the estimating equations for parameters θ (here $\theta = (\beta, \phi)$):

$$u(\theta) = \sum_i u_i(\theta) = 0$$

Denote the solution of these equations as $\hat{\theta}$ and the true θ as θ_0.

- Consistency: roughly speaking, when p is small, if $E(u(\theta_0)) \to 0$ when $n \to \infty$, then we can have $\hat{\theta} \to \theta_0$.
- Variance of $\hat{\theta}$. Under consistency, we can estimate the asymptotic variance of $\hat{\theta}$ by delta method.

By Delta method, we have

$$0 = u(\hat{\theta}) \approx u(\theta_0) + \dot{u}(\theta_0)(\hat{\theta} - \theta_0)$$

Thus, we have

$$\hat{\theta} - \theta_0 \approx -\dot{u}(\theta_0)^{-1} u(\theta_0)$$

Roughly speaking, we have

- Law of large numbers:

$$\frac{1}{n}\dot{u}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \dot{u}_i(\theta_0) \to E\left(\frac{1}{n}\sum_{i=1}^{n} \dot{u}_i(\theta_0)\right) = A$$

- CLT:

$$\frac{1}{\sqrt{n}}u(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i(\theta_0) \approx N(0, V)$$

Thus

$$\text{Var}(\hat{\theta}) \approx A^{-1}VA^{-T}/n$$

In practice, we estimate A and V by

$$\widehat{A} = \frac{1}{n}\sum_{i=1}^{n} \dot{u}_i(\hat{\theta})$$

and

$$\widehat{V} = \frac{1}{n}\sum_{i} u_i(\hat{\theta})u_i(\hat{\theta})^T$$

Next time: Examples of dispersed data and quasi-likelihood, mixed effect linear model