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 Violations of the variance assumptions in GLM

In earlier models, we typically have assumptions on the variance of $y_i \mid X_i$:

- In linear models, we assume $\mathrm{Var}(y_i) = \sigma^2$ (or more generally $\mathrm{Var}(y_i) = w_i\sigma^2$ with known w_i)
- In GLM with Binomial / Multinomial and Poisson distributions, we assume a fixed mean-variance relationship
- In practice, we can have over-dispersed/under-dispersed data or data with unequal variance.
- With wrong variance assumption but correct mean assumption (link function), we typically still get consistent point estimate $\hat{\beta}$ (though likely not the optimal one) and unreliable uncertainty quantification.

2 Over-dispersion

When we apply the standard GLM models assuming the data are Binomial or Poisson distributed to real data, it's common to see 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. $(y_i - \hat{\mu}_i)^2$ (as shown in R Data Example 6)

2.1 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 $\lambda_i \sim \mathrm{Gamma}(\mu_i, k_i)$. 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 $\gamma_i \equiv \gamma$.
- The ZIP / ZINB GLM can deal with over-dispersion caused by zero inflation

2.2 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

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

- If you treat y_i as a sum of Bernoulli variables $y_i = \sum_j Z_{ij}$ where $Z_{ij} \sim \mathrm{Bernoulli}(p_i)$, then randomness in p_i causes dependence among Z_{ij}.
- 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 $y_i \sim \text{Beta-binomial}(n_i, \mu_i, \theta)$ where $\mathbb{E}(y_i) = \mu_i$. The relation between μ_i and X_i are the same as we assumed for the standard binary GLM. For example:

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

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

3 Quasi-likelihood

The above solution replaces the exponential family distributions with a more complicated parametric distribution allowing an extra dispersion parameter in the variance. Another more general solution is to only assume a mean-variance relationship.

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 $\mathrm{Var}(y_i)$.
- Quasi-likelihood: we replace $\mathrm{Var}(y_i)$ by some other mean-variance relationship that we believe can better fit the data.
- Typically, the mean-variance relationship can 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 $a(\mu_i, \phi) = \mu_i + \phi\mu_i^2$ as in the Negative-Binomial distribution
- For grouped Binary data we can also assume $a(\mu_i, \phi) = \mu_i(n_i - \mu_i)\,(1 + (n_i - 1)\phi)$ as in the Beta-Binomial distribution

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 for β.
- $\mathrm{Var}(\hat{\beta})$ is affected by ϕ for any of the above mean-variance relationships.
- Including ϕ helps to get a correct uncertainty quantification of $\hat{\beta}$.

How to estimate ϕ? As we don't know the likelihood of the data, we only use moment conditions.

- 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

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

  We can solve ϕ by solving $X^2 = n - p$ (we use $n - p$ instead of n to correct for the degree of freedom in the estimated $\hat{\mu}_i$), which is

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

- For other forms of $a(\mu, \phi)$, we need to solve ϕ and β simultaneously from equations

$$\varphi_{1j}(\beta, \phi) = \frac{\partial L}{\partial \beta_j} = \sum_i \frac{(y_i - \mu_i)x_{ij}}{a(\mu_i, \phi)} \frac{1}{g'(\mu_i)} = 0 \tag{1}$$

$$\varphi_2(\beta, \phi) = \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{a(\mu_i, \phi)} - (n - p) = 0 \tag{2}$$

  - $\mathbb{E}[\varphi_{1j}(\beta, \phi)] = 0$ and $\mathbb{E}[\varphi_2(\beta, \phi)]/n \to 0$. Solutions $\hat{\beta}$ and $\hat{\phi}$ are called Z-estimators. Under proper regularity conditions, we can show that both $\hat{\beta}$ and $\hat{\phi}$ are consistent.

4 Estimating equations and Sandwich estimator

How to estimate the variance of $\hat{\beta}$ from the quasi-likelihood equations? And what if we do not even know the true form of the mean-variance relationship?

- The equations (2) is one type of estimating equations. In general, the estimating equations for parameters θ (here $\theta = (\beta, \phi)$ or $\theta = \beta$) have the form:

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

  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 $\hat{\theta} \to \theta_0$ (with some additional conditions).
  - Variance of $\hat{\theta}$. Under consistency, we can estimate the asymptotic variance of $\hat{\theta}$ by first-order Taylor expansion (see later).

- The score equations

$$u(\beta) = \sum_i \frac{(y_i - \mu_i)x_{ij}}{v^\star(\mu_i)} \frac{1}{g'(\mu_i)} = 0$$

  are valid estimating equations ($\mathbb{E}[u(\beta_0)] = 0$) as long as as the link function is correct. The response y_i does not need to follow the assumed exponential family distribution and $v^\star(\mu_i)$ does not need to be the correct form of variance.

- Even the simple $\sum_i (y_i - \mu_i)x_{ij} = 0$ are always valid estimating equations. The problem is that $\mathrm{sd}(\hat{\beta})$ may be large if samples have unequal variances.

Sandwich estimator of the asymptotic variances:
Let's now calculate the asymptotic variance of $\hat{\theta}$ for

$$\mu(\hat{\theta}) = 0$$

By first-order Taylor expansion, 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

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

In practice, we can estimate A and V by

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

and

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

- We use the sample variance to approximate V without knowing the distribution of the data
- The Sandwich estimator provides an estimate of the variance of $\hat{\beta}$ even when model assumption is violated.

Next time: Mixed effect linear models