STAT347: Generalized Linear Models
Lecture 10

Today's topics: Chapters 7.3-7.5

- Negative Binomial GLM
- Zero inflated models: ZIP, ZINB and hurdle models
- Revisit the example of the horseshoe crab dataset

1 Model for over-dispersed counts: Negative Binomial GLM

Think about the scenario $y_i \sim \text{Poisson}(\lambda_i)$ but $\log(\lambda_i) = X_i^T\beta + \epsilon_i$ indicating that X_i can not fully explain λ_i. Then

$$E(y_i) = E[E(y_i \mid \lambda_i)] = E(\lambda_i)$$

while

$$\text{Var}(y_i) = E[\text{Var}(y_i \mid \lambda_i)] + \text{Var}[E(y_i \mid \lambda_i)] = E(\lambda_i) + \text{Var}(\lambda_i) > E(y_i)$$

which show an over-dispersion of the distribution of y_i compared with a Poisson distribution.

- For example, we saw the over-dispersion issue in the horseshoe satellites dataset in Data Example 1 and homework 1, 1.22(a).
- Over-dispersion happens in Poisson and Binomial (Multinomial) GLM models as the variance is completely determined by the mean.
- There is no over-dispersion issue in linear models as linear models has an extra dispersion parameter.
- We will talk about general solutions for over-dispersion issues in later chapters.

For counts response, we can use a Negative binomial distribution to solve the over-dispersion issue.

Negative binomial distribution: $y \sim \text{Poisson}(\lambda)$ and $\lambda \sim \text{Gamma}(\mu, k)$ $[\mathbb{E}(\lambda) = \mu]$. The probability function of y is

$$f(y; \mu, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k$$

where $\gamma = 1/k$ is called a dispersion parameter.

- $\mathbb{E}(y) = \mu, \quad \text{Var}(y) = \mu + \gamma\mu^2$

- Negative Binomial distribution with fixed k belongs to the exponential family: $\theta = \log(\mu\gamma/(\mu\gamma+1))$ and $b(\theta) = -1/\gamma \log(\mu\gamma+1) = 1/\gamma \log(1-e^{\theta})$

Negative Binomial GLM:

- We assume $y_i \sim \text{NB}(\mu_i, k_i)$, with the link function $g(\mu_i) = X_i^T\beta$. Typically, we assume they share the same dispersion, so $\gamma_i = 1/k_i \equiv \gamma$ for all i.
- As an extension of Poisson GLM, a common link function is the log link: $g(\mu_i) = \log(\mu_i)$.
- When $g(\mu_i) = \log(\mu_i)$, The score equation for β is

$$\sum_i \frac{y_i - \mu_i}{\mu_i + \gamma\mu_i^2}\mu_i x_{ij} = \sum_i \frac{y_i - \mu_i}{1+\gamma\mu_i} x_{ij} = 0$$

- As $\mathbb{E}(\partial^2 L/\partial\beta_j\partial\gamma) = 0$, asymptotically $\hat{\beta}$ and $\hat{\gamma}$ are independent. Thus, the asymptotic variance of $\hat{\beta}$ would be the same no matter what γ is (Agresti book chapter 7.3.3).

$$\widehat{\text{Var}}(\hat{\beta}) = (X^T\hat{W}X)^{-1}$$

2 Models for zero-inflated counts

For a Poisson distribution $y \sim \text{Poisson}(\mu)$: $P(y=0) = e^{-\mu}$

For a Negative Binomial distribution $y \sim \text{NB}(\mu, k)$: $P(y=0) = \left(\frac{k}{\mu+k}\right)^k$

In practice, there may be way more 0 counts than what these distributions can allow. Example: y_i is the number of times going to a gym for the past week and there may be a substantial proportion who never exercise (you may see two modes in the distribution).

2.1 Zero-inflated Poisson / Negative Binomial (ZIP/ZINB) models

The ZIP model:

$$y_i \sim \begin{cases} 0 \quad \text{with probability } 1-\phi_i \\ \text{Poisson}(\lambda_i) \quad \text{with probability } \phi_i \end{cases}$$

We can interpret this as having a latent binary variable $Z_i \sim \text{Bernoulli}(\phi_i)$. If $z_i = 0$ then $y_i = 0$, and if $z_i = 1$ then y_i follows a Poisson distribution. For the GLM model, a common assumption for the links are:

$$\text{logit}(\phi_i) = X_{1i}^T\beta_1, \quad \log(\lambda_i) = X_{2i}^T\beta_2$$

- The mean is $E(y_i) = \phi_i\lambda_i$ and the variance is

$$\text{Var}(y_i) = \phi_i\lambda_i[1 + (1-\phi_i)\lambda_i] > E(y_i)$$

  So zero-inflation can also cause over-dispersion

- We may still see over-dispersion conditional on Z_i, then we can use a ZINB model where

$$y_i \sim \begin{cases} 0 & \text{with probability } 1-\phi_i \\ \text{NB}(\lambda_i, k) & \text{with probability } \phi_i \end{cases}$$

- We can use MLE to solve both the ZIP and ZINB model.

2.2 Hurdle model

The ZIP/ZINB model do not allow zero deflation. The Hurdle model separates the analysis of zero counts and positive counts.

Let

$$y_i' = \begin{cases} 0 & \text{if } y_i = 0 \\ 1 & \text{if } y_i > 0 \end{cases}$$

The Hurdle model assumes that $y_i' \sim \text{Bernoulli}(\pi_i)$ and $y_i \mid y_i > 0$ follows a truncated-at-zero Poisson ($\text{Poi}(\mu_i)$) / Negative Binomial ($\text{NB}(\mu_i, \gamma)$) distribution. Let the untruncated probability function be $f(y_i; \mu_i)$, then

$$P(y_i = k) = \pi_i \frac{f(k;\mu_i)}{1 - f(0;\mu_i)}, \quad \text{for } k \neq 0$$

$$P(y_i = 0) = 1 - \pi_i$$

For the GLM, we may assume

$$\text{logit}(\pi_i) = X_{1i}^T \beta_1, \quad \log(\mu_i) = X_{2i}^T \beta_2$$

- We can estimate β_i and β_2 separately using two separate likelihoods: $L(\beta_1, \beta_2) = L(\beta_1) + L(\beta_2)$
- There is zero deflation if $1 - \pi_i \leq f(0; \mu_i)$

3 Revisit the horseshoe crab data

Please see R notebook Example 6.