

# STAT347: Generalized Linear Models

## Lecture 10

Winter, 2023  
Jingshu Wang

# Today's topics:

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

# Over-dispersion in the Poisson model

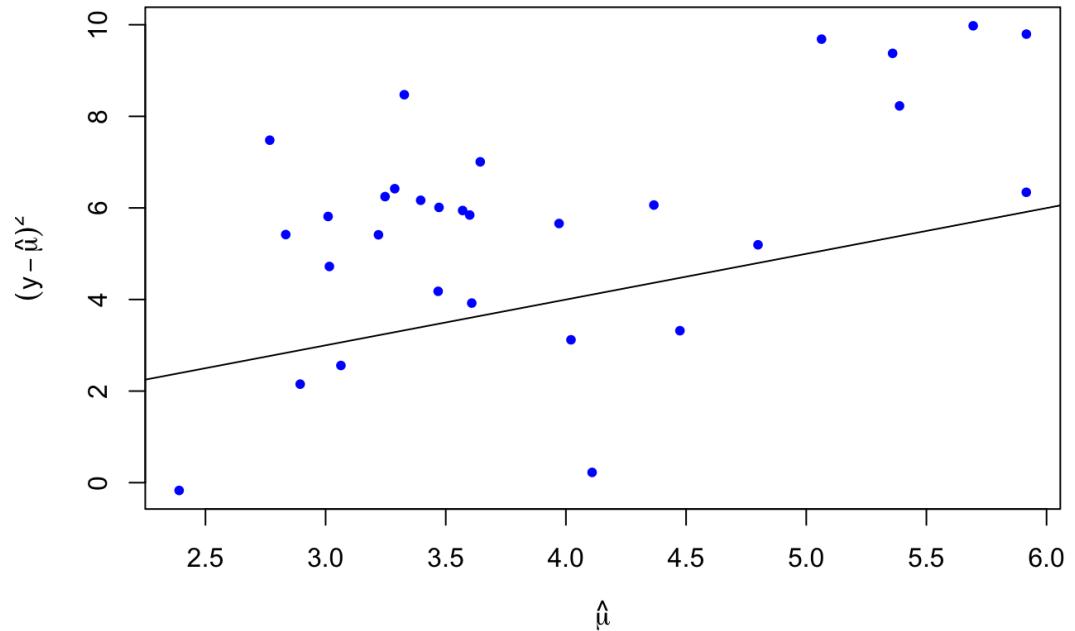
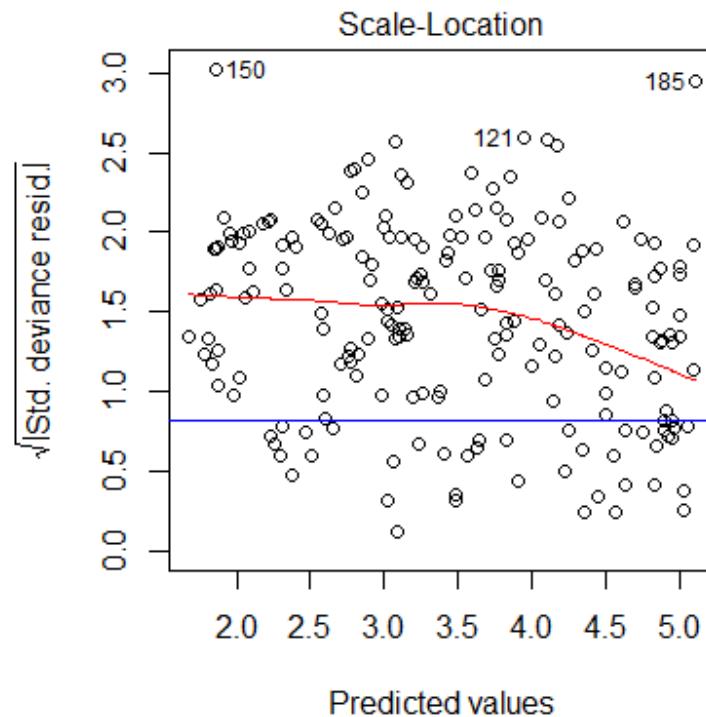
- Poisson regression assume that  $\text{Var}[y_i | X_i] = \mathbb{E}[y_i | X_i]$
- Over-dispersion: in practice, the counts  $y_i$  can be noisier than assumed in the Poisson distribution
- For instance, if  $\log(\lambda_i) = X_i^T \beta + \epsilon_i$  indicating that  $X_i$  can not fully explain  $\lambda_i$ . Then

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

while

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

# Over-dispersion examples



<https://stats.stackexchange.com/questions/331086/investigate-overdispersion-in-a-plot-for-a-poisson-regression>

<https://towardsdatascience.com/adjust-for-overdispersion-in-poisson-regression-4b1f52baa2f1>

# Over-dispersion in the Poisson model

- 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.

# Negative binomial distribution

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.

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

$$y_i \sim \text{NB}(\mu_i, k_i)$$

with the link function  $g(\mu_i) = X_i^T \beta$ .

- Typically, we assume that all samples share the same dispersion, so  $\gamma_i = \frac{1}{k_i} = \gamma$ .
- As an extension of the Poisson GLM, a common link for NB GLM is still the loglinear link:  $g(\mu_i) = \log(\mu_i)$
- Score equation for  $\beta$

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

# Negative binomial GLM

A bit about the inference:

- The hessian matrix has the term

$$\frac{\partial^2 L(\beta, \gamma; \mathbf{y})}{\partial \beta_j \partial \gamma} = - \sum_i \frac{(y_i - \mu_i)x_{ij}}{(1 + \gamma \mu_i)^2} \left( \frac{\partial \mu_i}{\partial \eta_i} \right).$$

Thus,  $E(\partial^2 L / \partial \beta_j \partial \gamma) = 0$  for each  $j$ , and  $\beta$  and  $\gamma$  are orthogonal parameters

- the asymptotic variance of  $\hat{\beta}$  would be the same no matter what  $\gamma$  is (Agresti book chapter 7.3.3)

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

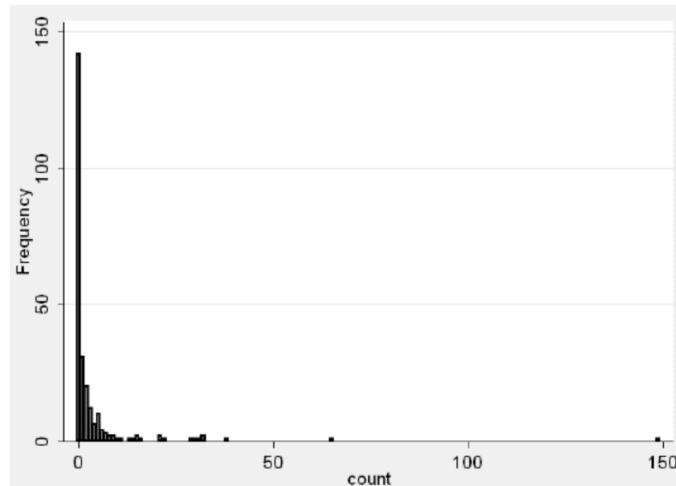
- $w_i = \mu_i / (1 + \gamma \mu_i)$

# 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



# Zero-inflated Poisson models

The ZIP model:

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

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

# Zero-inflated Negative Binomial models

- 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 still use MLE to solve both the ZIP and ZINB model
- The ZIP/ZINB model do not allow zero deflation.

# The Hurdle model

- 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  $y_i | y_i > 0$  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

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

# The Hurdle model

The joint likelihood function for the two-part hurdle model is

$$\ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \prod_{i=1}^n (1 - \pi_i)^{I(y_i=0)} \left[ \pi_i \frac{f(y_i; \mu_i)}{1 - f(0; \mu_i)} \right]^{1-I(y_i=0)},$$

where  $I(\cdot)$  is the indicator function. If  $(1 - \pi_i) > f(0; \mu_i)$  for every  $i$ , the model represents zero inflation. The log-likelihood separates into two terms,  $L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = L_1(\boldsymbol{\beta}_1) + L_2(\boldsymbol{\beta}_2)$ , where

$$L_1(\boldsymbol{\beta}_1) = \sum_{y_i=0} [\log(1 - \pi_i)] + \sum_{y_i>0} \log(\pi_i)$$

$$L_2(\boldsymbol{\beta}_2) = \sum_{y_i>0} \{ \log f(y_i; \exp(\mathbf{x}_{2i}\boldsymbol{\beta}_2)) - \log [1 - f(0; \exp(\mathbf{x}_{2i}\boldsymbol{\beta}_2))] \}$$

# Revisit the horseshoe crab data

- Check Example6 R notebook