

Over Dispersion

This occurs when $\text{var}(Y) > \mathbb{E}(Y)$.

Several sources of over dispersion:

- ▶ Correlated sampling: $Y = \sum_{i=1}^m Z_i$, where $Z_i \sim \text{Poisson}(\lambda_i)$ are correlated with $\text{corr}(Z_i, Z_j) = \rho_{ij} > 0$. Then

$$\mathbb{E}(Y) = \sum_{i=1}^m \lambda_i, \quad \text{var}(Y) = \sum_{i=1}^m \lambda_i + c > \mathbb{E}(Y)$$

- ▶ Clustering: $Y = U_1 + \dots + U_N$ where U_i are iid from some distribution taking integer values, and $N \sim \text{Poisson}(\lambda)$.
- ▶ Poisson-Gamma model: $Y|\mu \sim \text{Poisson}(\mu)$ and $\mu \sim \text{Gamma}(\alpha, \beta)$

$$Y \sim \text{X}_{\text{Poisson}}$$

Diagnostics with half-normal plot:

- ▶ Order the absolute value of residuals (Pearson or deviance residuals) without dispersion
- ▶ Plot $|r_{(i)}|$ (y coordinates) against $\Phi^{-1}\left(\frac{n+i+0.5}{2n+1.125}\right)$ (x coordinates), for $i = 1, \dots, n$
- ▶ Reference line is a straight line through origin with slope 1
- ▶ Linear deviation from the reference line indicates constant over-dispersion
- ▶ Empirical slope is roughly $\sqrt{\phi}$

Modeling Over-Dispersion

Two ways to model over dispersion:

- ▶ Poisson regression with a dispersion parameter
- ▶ Negative binomial regression

Poisson regression with a dispersion parameter:

In general, without knowing the mechanism that generates the over-dispersion, one can assume $\text{var}(Y) = \phi \mathbb{E}(Y)$ and estimate ϕ using

$$\hat{\phi} = \frac{\textcircled{G}}{n - p} = \frac{\sum (Y_i - \hat{\mu}_i)^2 / \hat{\mu}_i}{n - p}.$$

Then,

$$\begin{aligned}\tilde{\beta}_Q &= \hat{\beta}_{MLE} \\ \text{cov}(\tilde{\beta}_Q) &= \hat{\phi} \text{cov}(\hat{\beta}_{MLE})\end{aligned}$$

Deviance analysis with dispersion:

Assume we separate predictors into two sets $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. We want to test

$$H_0 : \beta_2 = \mathbf{0} \text{ vs } H_1 : \beta_2 \neq \mathbf{0}$$

- ▶ Model 1: $\eta = \mathbf{X}_1\beta_1$
- ▶ Model 2: $\eta = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$

Steps:

1. Calculate deviance D_1 from Model 1 (without over dispersion)
2. Calculate deviance D_2 from Model 2 (without over dispersion)
3. Estimate ϕ from the larger model (Model 2) by

$$\hat{\phi} = \underline{\underline{G/(n - p_1 - p_2)}}$$

4. Conduct F test

$$\frac{D_1 - D_2}{p_2 \hat{\phi}} \sim \underline{F(p_2, n - p_1 - p_2)}, \text{ under } H_0$$

Negative binomial regression:

- ▶ $Y \sim NB(r, p)$: # of successes before r failures (with success rate p)
- ▶ Mean $pr/(1-p)$, variance $pr/(1-p)^2$
- ▶ Reparameterization (mean $\mu > 0$ and dispersion parameter $\phi > 0$):

$$\mathbb{P}(Y = y) = \frac{\Gamma(y + \phi)}{\Gamma(y + 1)\Gamma(\phi)} \frac{\mu^y \phi^\phi}{(\mu + \phi)^{\phi+y}}$$

$$\begin{cases} \mu = \frac{pr}{1-p} \\ \phi = r \end{cases}$$

with mean μ and variance $\mu + \mu^2/\phi$.

- ▶ Negative-Binomial regression model can handle over dispersion:

$$\begin{cases} Y_i \sim NB(\mu_i, \phi) \\ \log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \end{cases}$$

Example: Wave Damage

For ship i , let n_{ij} be the number of months in the j th period, and Y_{ij} be the number of damages accordingly.

- ▶ Y_{ij} may be correlated across j
- ▶ This may cause over-dispersion

Zero-Inflated Poisson Regression

Sometimes, response may have excessive zeros

- ▶ State wildlife biologists want to model how many fish are being caught by visitors in a state park. Some visitors did not fish. Some visitors did fish but didn't catch any fish.
- ▶ In next-generation RNA sequencing study, gene expressions are measured by read counts. Some genes have no expression in a tissue sample, while others have but not detectable.

- ▶ Poisson or NB models tend to underestimate the number of zeros
- ▶ Need to model two processes separately:
 - ▶ one drives whether the value is always 0
 - ▶ one drives the value of potentially non-zero count
- ▶ Zero-inflated Poisson (ZIP) model

ZIP model for Y_i :

- ▶ Z_i is a latent binary variable that generates structural zeros

$\hookrightarrow \begin{matrix} Z_i = 0 & \text{true } 0 \\ Z_i = 1 & \text{pseudo } 0 \end{matrix}$

$$\mathbb{P}(Z_i = 0) = \pi_i \rightarrow p_r(\text{true } 0)$$

- ▶ The response satisfies

$$\left\{ \begin{array}{l} Y_i | (Z_i = 0) = 0 \\ Y_i | (Z_i = 1) \sim \text{Poisson}(\lambda_i) \end{array} \right.$$

- ▶ Consider two types of models

$\rightarrow \log(\lambda_i) = \mathbf{x}_i \boldsymbol{\beta}$ and $\text{logit}(\pi_i) = \mathbf{z}_i \boldsymbol{\gamma}$
Suitable for the fishing example

$\log(\lambda_i) = \mathbf{x}_i \boldsymbol{\beta}$ and $\text{logit}(\pi_i) = -\tau \mathbf{x}_i \boldsymbol{\beta} \ (\tau > 0)$
Suitable for the RNA-Seq example

Example: Fishing in the Park

It is of interest to investigate what factors are related to the number of fish caught in a state park.

camper	persons	child	count
0	1	0	0
1	1	0	0
0	1	0	0
1	2	1	0
0	1	0	1
1	4	2	0
...			

We assume whether a group fished or not depends on *persons*, and how many fish they got (if they fished) depends on *camper* and *child*.