STAT347: Generalized Linear Models
Lecture 11

Winter, 2023
Jingshu Wang

Today's topics:

- Beta-binomial GLM

Violation of the variance assumptions in GLM

In earlier models, we typically have assumptions on the variance of $y_i|X_i$

- Gaussian linear model: $\mathrm{Var}(y_i) = \sigma^2$
- GLM with Binomial / Multinomial / Poisson models: fixed mean-variance relationship

As we saw earlier, real data can have over-dispersion / under-dispersion or unequal variances, which violates these variance assumptions

- With wrong variance assumption but correct mean assumption (link function)
  - Typically still get consistent point estimate $\hat{\beta}$
  - Inference on $\hat{\beta}$ can be heavily impacted

Over-dispersion in the Poisson model

- Poisson regression assume that $\mathrm{Var}[y_i|X_i] = \mathbb{E}[y_i|X_i]$
- Over-dispersion: in practice, the counts y_i can be noisier than assumed in the Poisson distribution

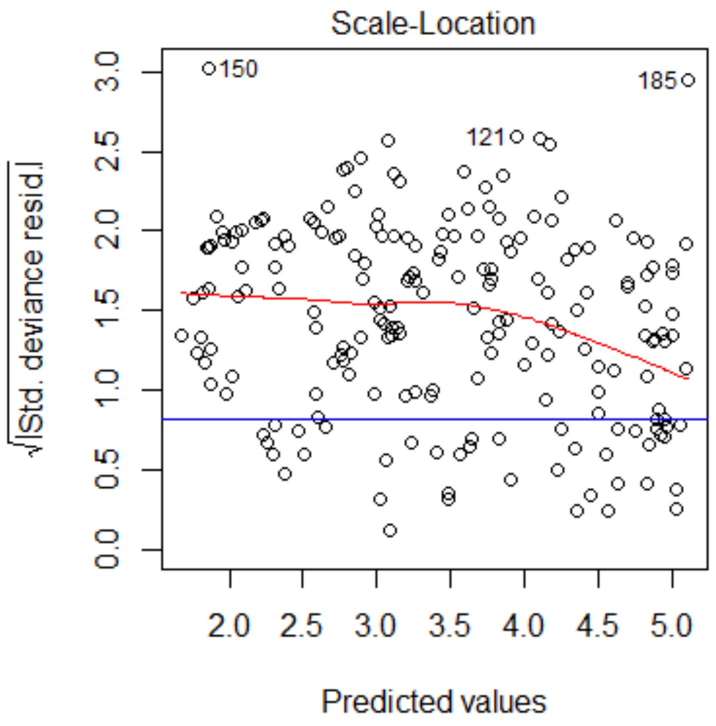
- For instance, if $\log(\lambda_i) = X_i^T \beta + \epsilon_i$ indicating that X_i can not fully explain λ_i. Then

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

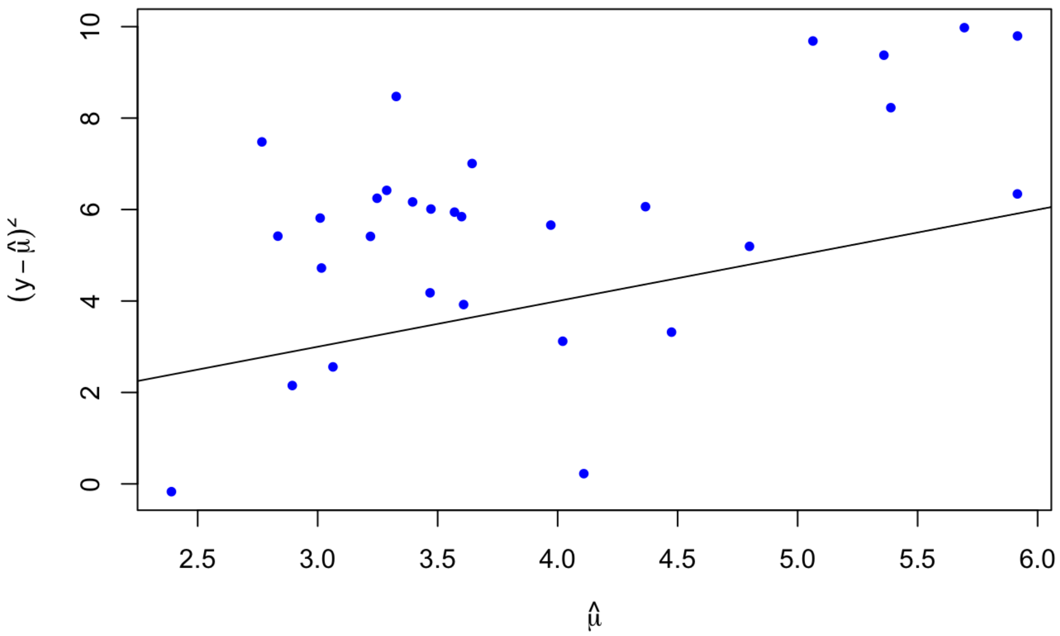
while

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

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

Variance inflation in binomial GLM

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.

What if the samples within each group are correlated?

- Analogous to the Poisson case, we can have the scenario

$$y_i \sim \text{Binomial}(n_i, p_i) \text{ but } \text{logit}(p_i) = X_i^T \beta + \epsilon_i$$

- Such a hierarchical model leads to variance inflation:

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

- If you treat y_i as a sum of Bernoulli variables $y_i = \sum_j Z_{ij}$ where $Z_{ij} \sim \text{Bernoulli}(p_i)$, then randomness in p_i causes dependence among Z_{ij}.

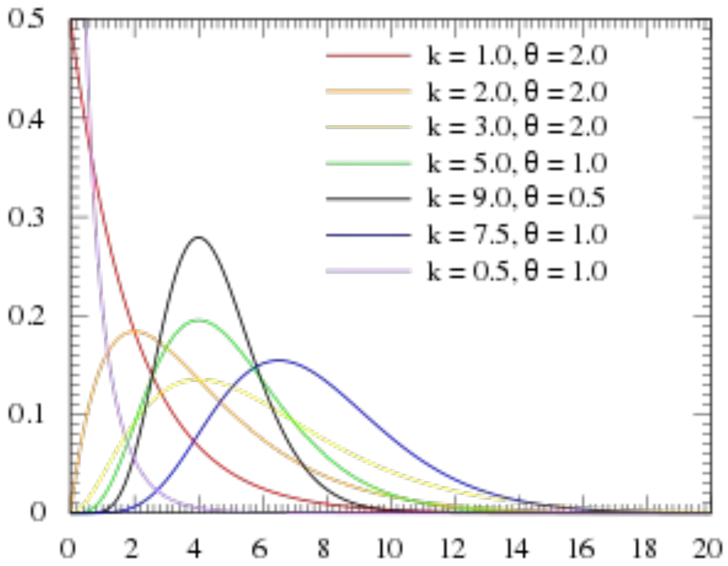
Negative binomial distribution

- Recall that we defined the negative binomial distribution for the over-dispersed counts

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

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

- It is defined as compound distribution (Gamma-Poisson mixture)



- Mean and variance of a Gamma distribution:

$$\mu = k\theta, \qquad \text{Var}(\lambda) = k\theta^2 = \frac{\mu^2}{k} = \gamma\mu^2$$

- For NB distribution

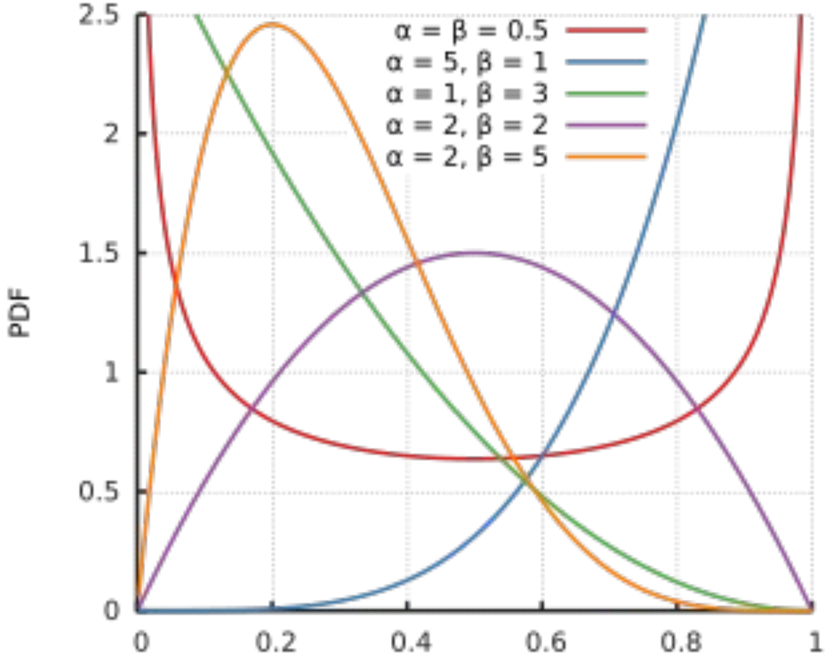
$$\mathbb{E}(y) = \mu, \quad \text{Var}(y) = \mu + \gamma\mu^2$$

Beta-binomial distribution

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

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

- Beta distribution



- Mean and variance of a Beta distribution:

$$\mu = \frac{\alpha}{\alpha + \beta},$$

$$\text{Var}(p) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \mu(1 - \mu)$$

- For Beta-binomial distribution distribution

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

where $\rho = 1/(\alpha + \beta + 1)$.

Beta-binomial GLM

- We assume that

$$y_i \sim \text{Beta-binomial}(n_i, \mu_i, \rho)$$

with the link function $g(\mu_i) = X_i^T \beta$. $\mathbb{E}(y_i) = n_i \mu_i$

  - As before, we assume that all samples share the same dispersion, so there is only one unknown dispersion parameter ρ.
  - A common link for Beta-binomial GLM is still the logit link:

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

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