# STAT347: Generalized Linear Models
# Lecture 11

Today's topics: Chapters 8

- Negative Binomial GLM and Beta-Binomial GLM

## 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 $\text{Var}(y_i) = \sigma^2$ (or more generally $\text{Var}(y_i) = w_i\sigma^2$ 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 $\text{Var}(y_i) > v^\star(y_i)$.
- 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: $y_i \sim \text{Poisson}(\lambda_i)$ and $\lambda_i \sim \text{Gamma}(\mu_i, k_i)$. Then $y_i \sim \text{NB}(\mu_i, k_i)$
- We have $E(y_i) = \mu_i$ and $\text{Var}(y_i) = \mu_i + \gamma_i\mu_i^2$ 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 $y_i \sim \text{Binomial}(n_i, p_i)$ but $\text{logit}(p_i) = X_i^T\beta + \epsilon_i$. We will then have

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

- 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 $Z_{ij}$.
- 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:

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

  and

  $$E(p) = \mu = \frac{\alpha}{\alpha+\beta}$$

  The Beta-binomial distribution then has the property that

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

  where $\rho = 1/(\alpha+\beta+1)$.
- Beta-binomial GLM:

  We assume the grouped data follows $y_i \sim$ Beta-binomial$(n_i, \mu_i, \rho)$ where $\mathbb{E}(y_i) = n_i\mu_i$. The relation between $\mu_i$ and $X_i$ are the same as we assumed for the standard binary GLM. For example:

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

  Both $\beta$ and $\rho$ are unknown but we can estimate using MLE.