# STAT347: Generalized Linear Models
## Lecture 1

Winter, 2024
Jingshu Wang

# Today’s topics: Agresti Chapter 1

- Two real data examples

- GLM concepts

# Two real data examples

- Example 1: Male Satellites for Female Horseshoe Crabs (Agresti section 1.5)
- Example 2: Election counts (Faraway Chapter 1)
  - Check Example1 R notebook

# Components of a generalized linear model (GLM)

Data points $(X_1, y_1), (X_2, y_2), \cdots, (X_n, y_n)$

- Random components: randomness in $y_i$ given $X_i$
  - Treat covariates $(X_1, \cdots, X_n)$ as fixed when performing statistical inference (same as in linear models)
  - Generalize $y_i$ from continuous real values to binary response, counts, categories, et. al.
  - We will start with assuming $y_i$ coming from an exponential family distribution.
    - Real valued response: Gaussian, Gamma (positive values)
    - Binary response: Bernoulli, Binomial
    - Counts: Poisson, Negative Binomial
    - Categorical response: Multinomial

# Components of a generalized linear model (GLM)

Data points $(X_1, y_1), (X_2, y_2), \cdots, (X_n, y_n)$

- Link function: how $\mathbb{E}(y_i)$ (or $\mathbb{E}(y_i|X_i)$) depends on $X_i$

$$g(\mathbb{E}(y_i)) = g(\mu_i) = X_i^T\beta \text{ where } \beta = (\beta_1, \cdots, \beta_p)^T \text{ and } X_i = (x_{i1}, \cdots, x_{ip})^T$$

- linear model: $g(\mu_i) = \mu_i$
- model for counts: $g(\mu_i) = \log(\mu_i)$.
- model for binary data: $g(\mu_i) = g(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$.

# Components of a GLM

Data points $(X_1, y_1), (X_2, y_2), \cdots, (X_n, y_n)$

- Linear predictor

$X\beta$ where $X = (X_1, X_2, \cdots, X_n)^T$ is the $n \times p$ model matrix.

- $X$ can include interactions, non-linear transformations of the observed covariates and the constant term
- avoid causal interpretations of the coefficients $\beta$ (read Chapter 1.2.3)

# GLM v.s. data transformation

- An alternative to GLM is to transform $y_i$ in some $h(y_i)$ a linear regression model of $h(y_i)$ on $X_i$
  - Commonly used in practice

Disadvantages:

- If $y_i$ are counts, usually take $h(y_i) = \log(y_i)$. How to deal with $y_i = 0$? How to transform binary or categorical data?
- need to find $h(\cdot)$ that can make a linear model reasonable as well as stabilizing the variance of $h(y_i)$.

Advantages:

- Easier to build models more complicated than a regression model in practice if we think the transformed data are approximately Gaussian.