STAT347: Generalized Linear Models
Lecture 1

Today's topics: Chapters 1, 4.1

- GLM concepts and examples
- Exponential family distributions for GLM

1 Components of a GLM

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

1. Random components: observations $(y_1, y_2, \cdots, y_n)$ follow some distribution family and are independent
    - Generalize y_i from continuous real values to binary response, counts, categories, et. al.
    - How to describe the distribution of y? We will start with assuming y_i coming from an exponential family distribution.
    - Treat the covariates $(X_1, \cdots, X_n)$ as fixed. For random X, build the model conditional on X.
2. Link function: $g(\mathbb{E}(y_i)) = g(\mu_i) = X_i^T\beta$ where $\beta = (\beta_1, \cdots, \beta_p)^T$ and $X_i = (x_{i1}, \cdots, x_{ip})^T$
    - linear model: $g(\mu_i) = \mu_i$
    - model for counts $g(\mu_i) = \log(\mu_i)$.
3. 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 β (read Chapter 1.2.3)

2 GLM v.s. data transformation

An alternative to GLM: transform y_i in some $h(y_i)$ and build and solve a linear model $h(y_i) = X_i^T\beta + \epsilon_i$.

- Sounds a reasonable approach, and is still commonly used now in various applications.
- 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? Also, the variance is not stabilized after transformation.
- Disadvantage of data transformation: need to find h that can make a linear model reasonable as well as stabilizing the variance. (read Chapter 1.1.6)
- Advantage of data transformation in practice: easier to build models more complicated than a regression model if we think the transformed data are approximately Gaussian.

3 Example: the horseshoe crab

Chapters 1.5.1 and 1.5.3

4 One-parameter exponential families

4.1 Definition

The observation y_i follows a one-parameter exponential family distribution and has the density $f(y_i;\theta_i)$ of the form ("density" here including the possibility of discrete atoms.)

$$f(y;\theta) = e^{y\theta - b(\theta)} f_0(y)$$

Terminologies:

- θ: natural or canonical parameters
- y: sufficient statistics
- $b(\theta)$: normalizing or cumulant function
- Chapter 4.1.1 Definition (4.1) is a more general family and includes dispersion parameters (such as the variance parameter for Gaussian)

4.2 Moment relationships

Take the first and second derivative respect to θ for both sides of the equation

$$e^{b(\theta)} = \int e^{y\theta} f_0(y) dy$$

We can derive:

$$\mu = \mathbb{E}(y) = b'(\theta); \quad V_\theta = \mathrm{Var}(y) = b''(\theta)$$

This indicates that:

$$\frac{\partial \mu}{\partial \theta} = \mathrm{Var}(y) > 0$$

thus the mapping from θ to μ is one to one increasing.

4.3 Some well-known one-parameter exponential families

1. Normal with variance 1
2. Bernoulli
3. Binomial
4. Poisson

4.4 The canonical link function in GLM

Assume

$$\theta_i = X_i^T \beta$$

(Why? Easier calculations)
As $\mu_i = b'(\theta_i)$, the link function will be

$$g(\cdot) = (b')^{-1}(\cdot)$$

which is called the canonical link.
Canonical link functions for Binomial, Poisson and Bernoulli distributions.

Next time: Chapter 4.2, ML estimation of GLM