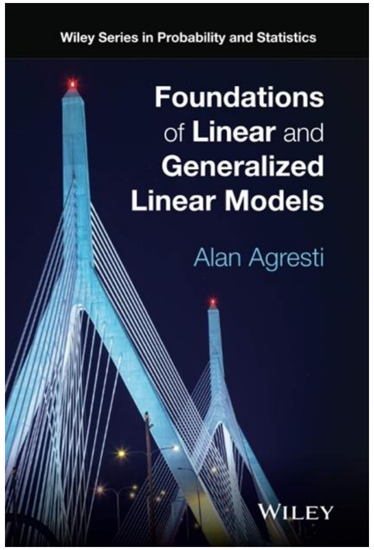
Generalized Linear Models

STAT34700, Winter 2025

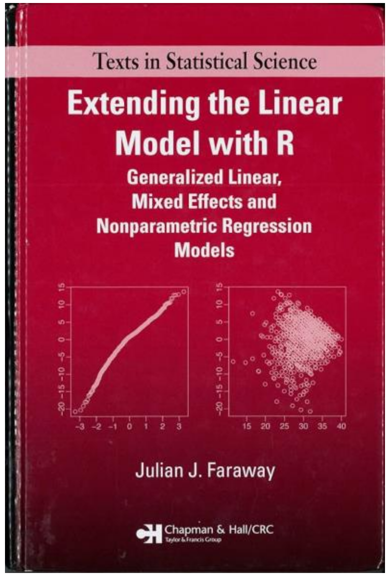
Jingshu Wang

Course logistics

- Focus of this course:
  - Basic GLM concepts, assumptions of different models and interpretation
  - GLM from both parametric and semi-parametric perspectives
  - Mathematical derivations of statistical estimation, inference and diagnosis
  - Using GLM and R software to analysis real datasets


Wiley Series in Probability and Statistics
Foundations of Linear and Generalized Linear Models
Alan Agresti
WILEY

- Textbooks
  - Foundations of Linear and Generalized Linear Models by Alan Agresti
  - Extending the Linear Model with R (second edition) by Julian J. Faraway
  - E-sources available on UChicago library website


Texts in Statistical Science
Extending the Linear Model with R
Generalized Linear, Mixed Effects and Nonparametric Regression Models
Julian J. Faraway
Chapman & Hall/CRC

Course logistics

- Homework
  - Four homework: due at 11:59 CST on Fridays of week 2/4/6/8 at midnight.
  - Homework needs to be submitted online through Gradescope.

- Office hours
  - Instructor OH: Tuesdays 4:00 pm – 5:00 pm in Jones 317
  - TAs OH: TBD (see updates on Canvas)

- Exams
  - Midterm: Thursday of week 6 (Feb 13) in class
  - Final: TBD

- Grading policy: Homework 20%, Midterm 40%, final 40%

Course logistics

- Use of AI tools
  - We encourage use of AI tools to help you learn course materials
  - We encourage use of AI tools to help you perform data analysis: data preprocessing, data visualization, guidance in choosing models and interpretation
  - We discourage use of AI tools to replace thinking and learning by yourself

  - AI tools can be used in your homework
  - AI tools are forbidden in exams

Lecture 1
Introduction to GLM concepts

Today’s topics:

- Review of Gaussian linear models
- Two real data examples
- GLM concepts
- Reading: Agresti Chapter 1, Faraway Chapters 1, 8.1

Gaussian linear model

Data points $(\boldsymbol{X}_1, y_1)$, …, $(\boldsymbol{X}_n, y_n)$

- Each $\boldsymbol{X}_i = (x_{i1}, \cdots, x_{ip})$

- Gaussian linear model:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

  - Can include intercept $x_{i1} = 1$

  - Relationship between $\mu_i = \mathbb{E}(y_i|\boldsymbol{X}_i)$ (also rewritten as $\mathbb{E}(y_i)$ treating $\boldsymbol{X}_i$ fixed) and $\boldsymbol{X}_i$
    - Linear relationship: $\mu_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
    - What if the relationship between μ_i and $\boldsymbol{X}_i$ is not linear?
      - Binary outcome, counts, …

  - Randomness of y_i: $y_i|\boldsymbol{X}_i$ follows a Gaussian distribution
    - $y_i|\boldsymbol{X}_i \sim N(\mu_i, \sigma^2)$ or equivalently $\varepsilon_i \sim N(0, \sigma^2)$
    - What if the distribution of y_i is not Gaussian?
    - What if the variance of $y_i|\boldsymbol{X}_i$ is not homoscedastic and depends on $\boldsymbol{X}_i$?

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 $(\boldsymbol{X}_1, y_1)$, …, $(\boldsymbol{X}_n, y_n)$ independent observations

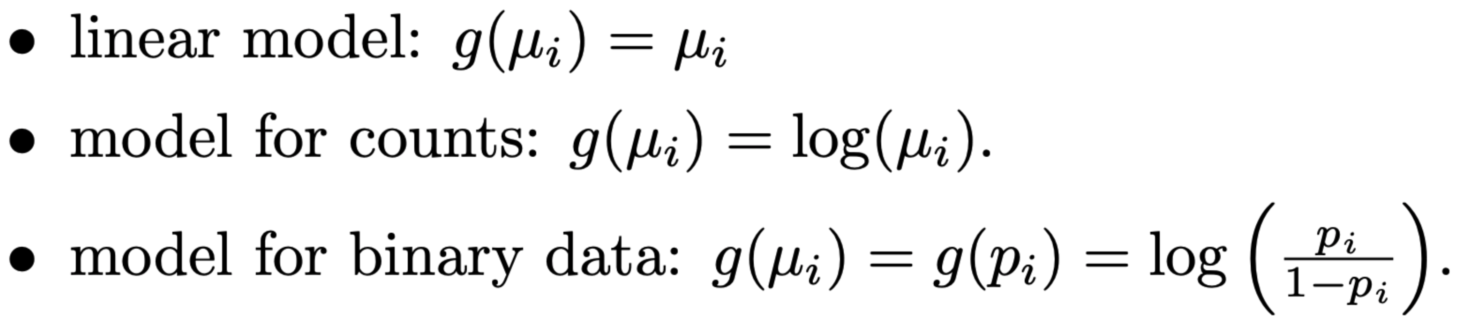
- Random components: randomness in y_i given $\boldsymbol{X}_i$
  - Treat covariates $(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$ as fixed when performing statistical inference (same as in linear models)
  - Generalize y_i from continuous real values to binary response, counts, categories, et. al.
  - 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 $(\boldsymbol{X}_1, y_1)$, …, $(\boldsymbol{X}_n, y_n)$

- Link function: how μ_i depends on $\boldsymbol{X}_i$
  - μ_i linearly depends on $\boldsymbol{X}_i$ after a pre-specified transformation

$$g(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\beta}$$



- 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 $(\boldsymbol{X}_1, y_1)$, …, $(\boldsymbol{X}_n, y_n)$

- Linear predictors: $\boldsymbol{X}_i = (x_{i1}, \cdots, x_{ip})$
  - $\boldsymbol{X}_i$ can include interactions, non-linear transformations of the observed covariates and the constant term
  - Having causal interpretations of the coefficients $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is challenging
    - More difficult than in linear regressions
    - β_j may not have a causal interpretations even if x_{ij} is completely randomized
      - Will discuss later in more details if we have time

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 the linear relationship reasonable as well as stabilizing the variance of $h(y_i)$.

  Advantages:

  - Easier to implement in practice
  - Especially useful for models that are more complicated than a regression model such as in factor models / PCA