STAT 528 - Advanced Regression Analysis II

Generalized Linear Models

Daniel J. Eck
Department of Statistics
University of Illinois

Agenda for today

- Course Syllabus and organization
- Introductory materials
- Course software and GitHub

Materials and organization

- Install or update R and RStudio
- Write homework assignments in RMarkdown and submit pdfs/html
- Submit homework in your personal GitHub repository within my GitHub STAT 528 organization

Computer resources

- The R Project for Statistical Computing: https://www.r-project.org/
- RStudio as an integrated development environment for R: https://www.rstudio.com/
- R Markdown: https://rmarkdown.rstudio.com/
- The tidyverse for data science: https://www.tidyverse.org/
- GitHub for version control: https://github.com/
- UIUC CS GitHub repo creator tool: https://wiki.illinois.edu/wiki/pages/viewpage.action?spaceKey=CSID&title=GitHub+repo+creator+tool

Course layout

- Our course will start with a detailed treatment of exponential family theory
- Motivation of GLMs from exponential family theory follows
- We will then discuss GLMs for binary and count responses as well as multinomial regression
- Next we will discuss data separation and GLM diagnostics
- Spring break!
- We will resume with contingency tables
- The remainder of the course will be devoted to a detailed treatment of more advanced regression topics:
  - linear mixed-effects models
  - generalized linear mixed-effects models and generalized estimating equations
  - aster models for life history analysis
  - multivariate regression and variance reduction via envelope methodology

Course layout (part II)

- The course will start with theory
- It will then be a combination of methodology and data analysis
- When we return from spring break, the course will be a mix of theory, methodology, and data analysis

Student Learning Outcomes

Upon successful completion of this course students will be able to conduct methodologically strong data analyses that can answer questions of scientific interest.

Students will gain written communication skills, will be able to present their data analyses in the form of a reproducible technical report, and will gain experience with data science workflow.

Background on topics

- Distributions
- Likelihoods
- Exponential families
- GLMs
- LMMs

There is more detail in the `introduction.pdf` notes. Some of this additional detail is useful for your first homework assignment.

Bernoulli and Binomial distributions

A random variable $Y \sim \text{Bernoulli}(p)$ has mass function

$$f(y) = \begin{cases} p & y = 1 \\ 1 - p & y = 0 \end{cases}$$

where $\text{E}(Y) = p$ and $\text{Var}(Y) = 1 - p$, and $0 < p < 1$ is a success probability.

A random variable $Y \sim \text{Binomial}(n, p)$ has mass function

$$f(y) = \binom{n}{y} p^y (1-p)^{1-y}, \qquad y = 0, 1, \ldots, n,$$

where $\text{E}(Y) = np$ and $\text{Var}(Y) = np(1-p)$. The Binomial distribution arises as a sum of Bernoulli trials.

Multinomial distribution

- n independent trials
- each trial results in one of c categories being observed
- probability vector $\mathbf{p} = (p_1, \ldots, p_c)$
- N_j be the total number of observations in category level

We will say

$$(N_1, \ldots, N_c) \sim \text{multinomial}(n, \mathbf{p})$$

with mass function

$$f(n_1, \ldots, n_c) = \binom{n}{n_1 \ldots n_c} p_1^{n_1} \cdots p_c^{n_c}, \qquad \sum_{j=1}^{c} n_j = n,$$

where:

- $\mathrm{E}(N_j) = np_j$
- $\mathrm{Var}(N_j) = np_j(1 - p_j)$
- $\mathrm{Cov}(N_j, N_k) = -np_jp_k$ where $j \neq k$.

Poisson distribution

A random variable $Y \sim \text{Poisson}(\mu)$, $\mu > 0$, has mass function

$$f(y) = \frac{\mu^y e^{-\mu}}{y!}, \qquad y = 0, 1, \ldots,$$

where $\text{E}(Y) = \mu$ and $\text{Var}(Y) = \mu$.

Likelihoods

- For a model with parameter θ, the **likelihood** $L(\theta)$ is the joint density of data at its observed values, as a function of θ.
- When data is iid the likelihood and log likelihood $l(\theta)$ will be

$$L(\theta) = \prod_{i=1}^{n} f_\theta(y_i), \qquad l(\theta) = \sum_{i=1}^{n} \log(f_\theta(y_1)).$$

- A maximum likelihood estimate (MLE) $\hat{\theta}$ maximizes $l(\theta)$ and $L(\theta)$. The estimate $\hat{\theta}$ is usually the unique solution of $\frac{\partial}{\partial\theta} l(\theta) = 0$.
- $\sqrt{n}(\hat{\theta} - \theta) \to N(0, I(\theta)^{-1})$ where

$$I(\theta) = -\mathrm{E}\left(\frac{\partial^2 l(\theta)}{\partial\theta^2}\right).$$

Exponential family

An *exponential family of distributions* is a parametric statistical model having log likelihood that takes the form

$$l(\theta) = \langle y, \theta \rangle - c(\theta), \tag{1}$$

where:

- y is a vector statistic
- θ is a vector parameter
- $\langle y, \theta \rangle$ is the usual inner product
- $c(\theta)$ is the cumulant function.

We have:

$$\mathrm{E}_\theta(Y) = \nabla c(\theta),$$
$$\mathrm{Var}_\theta(Y) = \nabla^2 c(\theta).$$

Examples: Binomial, Multinomial, Poisson, normal, etc.

GLM

Recall in LM:

$$y = x^T\beta + \varepsilon, \qquad \mathsf{E}(y|x) = x^T\beta$$

Generalized linear models (GLM) are extensions to the above where in GLM:

$$\mathsf{E}_\theta(y_i|x_i) = g(x_i^T\beta)$$

which implies that we can write

$$g^{-1}\left(\mathsf{E}_\theta(y_i|x_i)\right) = x_i^T\beta.$$

These models have very nice statistical properties when the underlying model is an exponential family.

LMM

The basic LMM takes the form

$$Y = X\beta + Zb + \varepsilon \qquad \text{or} \qquad Y \mid b \sim N(X\beta + Zb, \sigma^2 I),$$

where:

- Y is the response vector
- X is a fixed-effects model matrix
- β is a fixed-effects coefficient vector
- Z is a model matrix of random-effects
- b is a vector of random effects
- σ^2 is the variance of the error distribution.

If we further assume that $b \sim N(0, \sigma^2 D)$ then the unconditional response Y is distributed as

$$Y \sim N(X\beta, \sigma^2(I + ZDZ^T)).$$

Multivariate regression model

The basic multivariate regression model takes the form

$$Y = \alpha + \beta X + \varepsilon,$$

where:

- Y is the response vector
- α is an intercept vector
- X is a predictor vector
- β is a coefficient matrix
- $\varepsilon \sim N(0, \Sigma)$ where $\Sigma > 0$.