STAT 545/GS01 1233 193 5528: Generalized Linear Models and Categorical Data Analysis (Part II) 1: GLM Introduction

Ziyi Li

Department of Biostatistics
The University of Texas MD Anderson Cancer Center

ZLi16@mdanderson.org

Fall 2022

Overview of Part II of this class

- Oct 14, 2022 to December 2, 2022
- There will be three homework and a final exam on the last day of class
- Selected sections from Chapters 4-8, 11-13 of Alan Agresti book
  - Regression model for binary data
  - Regression model for ordinal data
  - Regression model for counts data
  - Marginal regression models for longitudinal categorical data
  - Conditional regression models for longitudinal categorical data
- Common theme: studying the association between covariates (explanatory variables) and the outcome (response variable; categorical data)

Tentative schedule

	Date	Content		
1	10/14/2022	GLM:intro		
2	10/17/2022	GLM:intro		
3	10/19/2022	GLM:intro/est		Assign 1st HW
4	10/21/2022	GLM: est	conflict with NIH review meeting	
5	10/24/2022	logistic		
6	10/26/2022	logistic		
7	10/28/2022	logistic		1st hw due
8	10/31/2022	logistic		
9	11/2/2022	ordinal data		Assign 2nd HW
10	11/4/2022	ordinal data		
11	11/7/2022	ordinal data		
12	11/9/2022	ordinal data		
13	11/11/2022	Poisson		
14	11/14/2022	Poisson		2nd hw due
15	11/16/2022	matched pairs		
16	11/18/2022	GEE		Assign 3rd HW
17	11/21/2022	GEE		
	11/23/2022	Thanksgiving		
	11/25/2022	Thanksgiving		
18	11/28/2022	GLMM		
19	11/30/2022	GLMM		3rd HW due
20	12/2/2022	Exam	50 min	

Introduction

- Outcome: $Y_i, i = 1, \cdots, n$, independent with different means $E(Y_i) = \mu_i$.
- Predictors/Covariates: $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$; let $x_{i1} = 1$ for the intercept term unless otherwise noted.
- Goal: Assess the relationship between Y_i and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$

Introduction

- Outcome: $Y_i, i = 1, \cdots, n$, independent with different means $E(Y_i) = \mu_i$.
- Predictors/Covariates: $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$; let $x_{i1} = 1$ for the intercept term unless otherwise noted.
- Goal: Assess the relationship between Y_i and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$

$$Y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i = \mu_i + \epsilon_i, \text{ where } \mu_i = E(Y|\boldsymbol{x}_i).$$

$$Y_i \sim N(\mu_i, \sigma^2) \ , \ \text{with } \mu_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

- A typical linear model includes a systematic component and a random component, which is a decomposition of the data into signal (explained by the model) and noise (unexplained variation). The random component follows a normal distribution.

From LM to GLM

The two components (that we are going to relax) are

- **Random component**: The response variable $Y|\boldsymbol{x}$ is continuous and normally distributed with mean $\mu = \mu(X) = E(Y|X)$.

- **Link**: between the random and covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$: $\mu(\boldsymbol{x}_i) = \boldsymbol{x}^T\beta$

From LM to GLM

The two components (that we are going to relax) are

- **Random component**: The response variable $Y|\boldsymbol{x}$ is continuous and normally distributed with mean $\mu = \mu(X) = E(Y|X)$.

  Replacing the normal distribution with the exponential family distribution.

  Allowing for non-constant variance of Y_i.

- **Link**: between the random and covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$: $\mu(\boldsymbol{x}_i) = \boldsymbol{x}^T\beta$

  Impose a link function to connect μ and the systematic component: $g(\mu_i) = \boldsymbol{x}^T\beta$

GLM: Three Components

1. **Random component:**

$$\begin{aligned} Y_i &\sim f(y_i; \theta_i, \phi) \\ &= \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \end{aligned}$$

2. **Systematic component:**

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

where η_i is the linear predictor and $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is the coefficient vector.

3. **Link Function:**

$$\eta_i = g(\mu_i)$$

where $g(\cdot)$ is a smooth monotonic function, known as the link function.

Creating A Model for Binary Outcome: Logistic Regression

1. Data example from 1998 Behavioral Risk Factors Social Survey. N=30,000+ women contacted by mail
2. Goal: study the association between demographic factors and women's use of oral contraceptives in the US
3. Outcome: 1 = oral contraceptives used, 0 = never used
4. Covariates: age, race, years of education, marital status
5. Unit of analysis: individual women (sample size = 30,000+)
6. The covariates do not determine the use or non-use of the oral contraceptives. Among all the women with a given combination of covariates, the binary outcome follows a probability distribution, and that distribution can be reasonably assumed to be Bernoulli. The Bernoulli probability can never be 0 or 1.

Creating A Model for Binary Outcome: Logistic Regression

$$Y_i \sim \text{Bernoulli}(p_i) \text{ with } g(p_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

$$L(\beta_0, \beta_1, \beta_2) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \quad \text{likelihood/density}$$

1. Linear model for continuous outcome with normal distribution at its core. Model for binary outcome: Bernoulli distribution (a special case of binomial distribution)
2. Unlike the normal distribution, which has two parameters and we only model the mean, the Bernoulli distribution has one parameter and we explain it with covariates. NOTE: $E(Y_i) = p_i$
3. What are the systematic and random component?
4. The $g(.)$ can be any monotone increasing function that maps $(0, 1)$ to $(-\infty, +\infty)$. Common choices:

$$\textit{logit}(p) = \textit{log}\left(\frac{p}{1-p}\right) \text{ and } \textit{probit}(p) = \Phi^{-1}(p)$$

Link Function

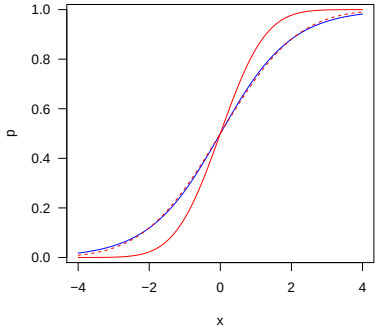
1. We call $g(.)$ a link function, which maps $(0, 1)$ (the scale of the mean outcome) to $(-\infty, +\infty)$ (the scale of linear combination of covariates).
2. We call $g^{-1}(.)$ an inverse link function, which maps $(-\infty, +\infty)$ (the scale of linear combination of covariates) to $(0, 1)$ (the scale of the mean outcome).
3. Any continuous, strictly increasing function that maps $(-\infty, +\infty)$ to $(0, 1)$ can serve as the inverse of a link function. The obvious choice is the CDF of a continuous distribution. Therefore, the link function is just the inverse of the CDF of a chosen continuous distribution.

Examples of Link Function

- Identity link: $\eta_i = g(\mu_i) = \mu_i$.
- Logit link: $\eta_i = g(\mu_i) = \log \frac{\mu_i}{1-\mu_i}$.
- Probit link: $\eta_i = g(\mu_i) = \Phi^{-1}(\mu_i)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable.
- Complementary log-log link: $\eta_i = g(\mu_i) = \log(-\log(1-\mu_i))$.
- Log link: $\eta_i = g(\mu_i) = \log(\mu_i)$.
- Inverse link: $\eta_i = g(\mu_i) = \frac{1}{\mu_i}$.
- **Canonial link:** When $\eta_i = g(\mu_i) = \theta_i$ where θ_i is the natural parameter, $g(\cdot)$ is known as the canonical link, which is specific to the distribution of Y_i.

Logit and Probit Functions

$$logit(p) = log\left(\frac{p}{1-p}\right) \quad \text{and} \quad expit(x) = \frac{\exp(x)}{1+exp(x)} = \frac{1}{1+exp(-x)}$$



Logit and Probit Functions

$$logit(p) = log\left(\frac{p}{1-p}\right) \quad \text{and} \quad expit(x) = \frac{\exp(x)}{1+exp(x)} = \frac{1}{1+exp(-x)}.$$

$$p_i = expit(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})}{1+exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})}$$

1. logit(.) is the link function, and expit(.) is the inverse link function.
2. The expit(.) function is also called a **logistic function**. Since we use logistic function to establish the connection between covariates and the Bernoulli distribution, this is called **logistic regression**.
3. Logistic function (blue) and inverse probit function (red) are in similar shape
4. Logistic function (blue) and $N(0, 1.7^2)$ CDF (pink) are very close
5. Logistic regression is more widely used because its parameters have better interpretation (to be elaborated later)

Example: Logistic Regression

1. **Random Components:** Probability mass function:

$$f(y_i; p_i) = p_i^{y_i}(1-p_i)^{1-y_i} = \exp\left\{y_i \log\frac{p_i}{1-p_i} + \log(1-p_i)\right\}$$

where $p_i = \Pr(Y_i = 1)$, $\theta_i = \log\frac{p_i}{1-p_i}$, and $a(\phi) = 1$.

2. **Systematic Components:**

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

3. **Canonical Link Function:** Logit link

$$\eta_i = g(p_i) = \theta_i = \log\frac{p_i}{1-p_i}.$$

Combining 2-3, we have the standard logistic regression model

$$\log\frac{p_i}{1-p_i} = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

Creating A Model for Binomial Outcome: Data Example

1. Data example from a clinical trial comparing 5 drug doses, 100mg to 500mg
2. Study design: 5-10 patients assigned to each dose
3. Outcome: number of patients responding to the drug at a given dose level. Range: $0 - m$ where m is the number of patients assigned to that dose
4. Covariate: $x = 100, 200, 300, 400, 500$.
5. Unit of analysis: dose levels (sample size is 5)
6. Study goal: how the response rate change with dose level?
7. The response rate is a probability p_j associated with each dose level j, where $j = 1, 2, 3, 4, 5$. The outcome Y_j is a count, which has binomial distribution given p_j and m_j, the number of patients at dose level j

Creating A Model for Binomial Outcome: Binomial Regression

$$Y_j \sim \text{Binomial}(p_j, m_j) \text{ with } g(p_i) = \theta_i = \beta_0 + \beta_1 X_i$$

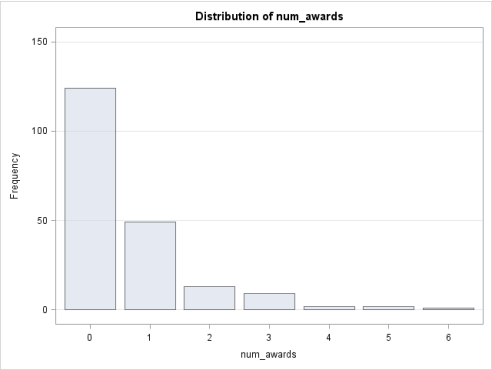
$$L(\beta_0, \beta_1) = \prod_{j=1}^{5} \binom{m_j}{y_j} p_j^{y_j} (1 - p_j)^{m_j - y_j} \quad \text{likelihood/density}$$

1. This binomial regression is equivalent to a logistic regression with the patient as the unit of analysis, and a sample size of $m_1 + m_2 + m_3 + m_4 + m_5$ patients. The likelihood is the same because the probability of drug response is the same for all patients assigned to a given dose
2. So we will only deal with logistic regression (Bernoulli distribution)
3. What are the systematic and random component? What is the canonical link?

Creating A Model for Counts Outcome: Data Example

1. Data example from a study on the association between the number of awards earned by students at one high school (4 years) and the students' performance in math and the type of program (e.g., vocational, general or academic) in which students were enrolled.
2. Unit of analysis: students
3. Outcome: number of awards each student received, ranging from 0, 1, 2, ...
4. Covariates: type of program (e.g., vocational, general or academic) and a score quantifying math performance

Data Example for Poisson Regression



Creating A Model for Counts Outcome: Poisson Regression

1. The two covariates are related to the outcome, but do not have a deterministic relationship
2. Consider Poisson distribution
3. The Poisson distribution has one parameter and we explain it with covariates. NOTE: $E(Y_i) = \lambda_i$

Creating A Model for Counts Outcome: Poisson Regression

$$Y_i \sim \text{Poisson}(\mu_i) \text{ with } \mu_i = exp(\beta_i) = exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})$$

$$L(\beta_0, \beta_1, \beta_2) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i} \quad \text{likelihood/density}$$

1. What are the systematic and random component? What is the canonical link?

2. We use *log*(.) link function. Therefore, the inverse link is *exp*(.). Any other strictly monotone increasing function that maps $(0, \infty)$ to $(-\infty, +\infty)$ can also serve as a link function, but *log*(.) link results in some convenient interpretation (to be elaborated later)

Mean and Variance Functions

The general likelihood result indicates that, if the model is correct and the likelihood satisfies the exponential family, then

$$E\left(\frac{\partial \log L_i}{\partial \theta_i}\right) = 0 \text{ and } -E\left(\frac{\partial^2 \log L_i}{\partial \theta_i^2}\right) = E\left\{\left(\frac{\partial \log L_i}{\partial \theta_i}\right)^2\right\}$$

The log-likelihood of all the data is $\sum_{i=1}^{n} \log L_i(\theta_i, \phi)$

$$\log L_i \equiv \log L_i(\theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)$$

where θ_i can be thought of as a subject-specific parameter (a deterministic function of covariates; no randomness).
Calculate the first and second derivatives with respect to θ_i:

$$\frac{\partial \log L_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} \text{ and } \frac{\partial^2 \log L_i}{\partial \theta_i^2} = -\frac{b''(\theta_i)}{a(\phi)}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ denote the first and second derivatives of $b(\theta_i)$ evaluated at θ_i.

Mean and Variance Functions

The first equality implies that

$$E\left(\frac{Y_i - b'(\theta_i)}{a(\phi)}\right) = 0$$

Therefore,

$$\mu_i = E(Y_i) = b'(\theta_i).$$

This is the mean expressed as a function of the natural parameter.

The second equality implies that

$$\frac{b''(\theta_i)}{a(\phi)} = E\left(-\frac{\partial^2 L_i}{\partial \theta_i^2}\right) = E\left\{\frac{[Y_i - b'(\theta_i)]^2}{a(\phi)^2}\right\} = \frac{var(Y_i)}{a(\phi)^2}$$

Therefore,

$$var(Y_i) = b''(\theta_i)a(\phi).$$

This is the variance expressed as a function of the natural parameter and dispersion parameter.

Mean and Variance Functions of Linear, Logistic and Poisson

Linear: $b(\theta_i) = \theta_i^2/2$ and $a(\phi) = \phi$ ($\phi = \sigma^2$)

$$E(Y_i) = b^{'}(\theta_i) = \theta_i \text{ and } var(Y_i) = b^{''}(\theta_i)a(\phi) = \phi$$

Logistic: $b(\theta_i) = log[1 + exp(\theta_i)]$ and $a(\phi) = 1$

$$E(Y_i) = b^{'}(\theta_i) = \frac{exp(\theta_i)}{1 + exp(\theta_i)} = expit(\theta_i) = \mu_i$$
$$var(Y_i) = b^{''}(\theta_i)a(\phi) = \mu_i(1 - \mu_i)$$

Poisson: $b(\theta_i) = exp(\theta_i)$ and $a(\phi) = 1$

$$E(Y_i) = b^{'}(\theta_i) = exp(\theta_i)$$
$$var(Y_i) = b^{''}(\theta_i)a(\phi) = exp(\theta_i)$$

How to solve GLM: Likelihood Equations

- Given GLM, our goal is to perform estimation and inference for the regression coefficients, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$. We consider the maximum likelihood approach.
- The observed data log-likelihood is

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_i \ell_i(\boldsymbol{\beta}; y_i) = \sum_i \{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\},$$

  where θ_i's are the functions of $\boldsymbol{\beta}$.
- Taking the first derivative w.r.t $\boldsymbol{\beta}$, we have the likelihood equations (score functions):

$$U(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = 0$$

How to solve GLM: Likelihood Equations

- Since ℓ is not an explicit function of $\boldsymbol{\beta}$, we need to obtain the likelihood equations using the chain rule.
- Recall $\mu_i = b'(\theta_i)$, $\eta_i = g(\mu_i)$, $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$.
- Then, we have

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_j} \frac{\partial \theta_i}{\partial \mu_j} \frac{\partial \mu_i}{\partial \eta_j} \frac{\partial \eta_i}{\partial \beta_j}$$

- In order to obtain $\dfrac{\partial \ell_i}{\partial \beta_j}$, we need to find out each individual term on the right-hand side of the equation.

How to solve GLM: Likelihood Equations

- We want to obtain

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_j}\frac{\partial \theta_i}{\partial \mu_j}\frac{\partial \mu_i}{\partial \eta_j}\frac{\partial \eta_i}{\partial \beta_j}$$

- We have $\mu_i = b'(\theta_i)$, $Var(Y_i) = b''(\theta_i)a(\phi)$ and $\eta_i = g(\mu_i)$, then
  (1) $\frac{\partial \ell_i}{\partial \theta_j} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$
  (2) $\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = Var(Y_i)/a(\phi)$
  (3) $\frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i)$
  (4) $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$
- Therefore

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)}\frac{a(\phi)}{Var(Y_i)}\frac{1}{g'(\mu_i)}x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)}\frac{1}{g'(\mu_i)}$$

How to solve GLM: Likelihood Equations

- The score equations are

$$U_j(\beta) = \sum_i \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \frac{1}{g'(\mu_i)} = 0, j = 1, 2, \cdots, p$$

  or in a matrix form

$$U(\beta) = \boldsymbol{X}^T \boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu}) = 0$$

  where $\Delta = diag\{\frac{1}{Var(Y_i)g'(\mu_i)}\}$, $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^T$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^T$
- $U(\beta)$ depends on the link function $g(\cdot)$. Different link functions give rise to different likelihood equations and hence different parameter estimates.
- Note that μ_i is a function of β.

How to solve GLM: Likelihood Equations

- When the canonical link function is used, we have $\eta_i = g(\mu_i) = \theta_i$ and hence

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i}$$

- Then the score equations reduce to

$$U_j(\boldsymbol{\beta}) = \frac{\partial \ell_i}{\partial \theta_j}\frac{\partial \theta_i}{\partial \mu_j}\frac{\partial \mu_i}{\partial \eta_j}\frac{\partial \eta_i}{\partial \beta_j} = \sum_i \frac{(y_i - \mu_i)x_{ij}}{a(\phi)} = 0, j = 1, 2, \cdots, p$$

- Since $a(\phi)$ is a constant, $U(\boldsymbol{\beta})$ can further reduce to

$$\sum_i (y_i - \mu_i)x_{ij} = 0, j = 1, 2, \cdots, p$$

or in a matrix form

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}) = 0$$

Summary of GLM

1. This chapter is an overview of GLM regression
2. We discussed the motivating ideas behind model development for categorical data
3. The typical GLM models (linear, logistic/binomial, Poisson) belong to natural exponential family of distributions, with canonical link
4. Unified derivation of mean and variance functions
5. Unified derivation of likelihood equations, asymptotic covariance matrix, model diagnosis statistics (to be covered in later chapters)
6. **Generalized linear model** (NOT General linear model): linear combination of covariates linked to the mean of the outcome variable with a twist (possibly nonlinear link function)
7. Next, we will discuss about estimation and model checking.