Part I: Generalized Linear Model

Outline

- Motivating examples
- Exponential family distributions
- Generalized linear model basics
- Logistic regression
- Multinomial regression
- Poisson regression
- Contingency table

Example I

The Kyphosis data consist of measurements on 81 children following corrective spinal surgery. The binary response variable, Kyphosis, indicates the presence or absence of a postoperative deforming. The three covariates are, Age of the child in month, Number of the vertebrae involved in the operation, and the Start of the range of the vertebrae involved. The first five observations are shown below.

	Kyphosis	Age	Number	Start
1	absent	71	3	5
2	absent	158	3	14
3	present	128	4	5
4	absent	2	5	1
5	absent	1	4	15

Questions of interest are:

- How do the three explanatory variables relate to the response?
- Can they be used to screen the patients prior to the operation?

Due to the binary nature of the response, it's not reasonable to model it as a linear function of the covariates. A more appropriate model would be the logistic regression:

$$Y_i \sim Bernoulli(\pi_i),$$

and

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{X}_i\boldsymbol{\beta}.$$

Example II

In a study of motor vehicle safety, 150 men and 150 women were interviewed to rate how important air conditioning and power steering were to them when they were buying a car.

Table: Importance Rating

		Response			
Sex	Age	Unimportant	Import	Very Import	Total
Women	18-23	26 (58%)	12 (27%)	7 (16%)	45
	24-40	9 (20%)	21 (47%)	15 (33%)	45
	> 40	5 (8%)	14 (23%)	41 (68%)	60
Men	18-23	40 (62%)	17 (26%)	8 (12%)	65
	24-40	17 (39%)	15 (34%)	12 (27%)	44
	> 40	8 (20%)	15 (37%)	18 (44%)	41
Total		105	94	101	300

Question of interest:

- How are sex and age related to the car preference?

The response of each subject is the preference level, which is categorical and ordinal. One plausible model would be

$$Y_i \sim multinomial(\pi_{i1}, \pi_{i2}, \pi_{i3}),$$

and

$$\log\left(\frac{\pi_{i1}}{\pi_{i2}+\pi_{i3}}\right) = \mathbf{X}_i\boldsymbol{\beta}_1;\ \log\left(\frac{\pi_{i1}+\pi_{i2}}{\pi_{i3}}\right) = \mathbf{X}_i\boldsymbol{\beta}_2.$$

We will discuss in detail when we introduce the multinomial logistic regression.

Example III

The example concerns a type of damage caused by waves to the forward section of certain cargo-carrying vessels. For the purpose of setting standards for hull construction we need to know the risk of damage associated with the three classifying factors shown below:

- Ship Type: A-E;
- Year of Construction: 1960-64, 65-69, 70-74, 75-79;
- Period of operation: 1960-74, 75-79.

Two other variables are Aggregated months of service and Number of damage accidents.

ship	year	period	month	damage
A	60-64	60-74	127	0
A	60-64	75-79	63	0
A	65-69	60-74	1095	3
A	65-69	75-79	1095	4
A	70-74	60-74	1512	6

Question of interest:

- How does the number of damage accidents depend on the other variables?

The response is count-valued, and may follow a Poisson distribution. A more appropriate model would be

$$Y_i \sim Poisson(\lambda_i),$$

and

$$\log \lambda_i = \log m_i + \mathbf{X}_i \boldsymbol{\beta}$$

where m_i is the number of months of service.
This is a Poisson regression model with offset.

Exponential Family

- Exponential family is a large family of distributions
- Many well known distributions are in the exponential family (e.g., normal, exponential, Poisson, Bernoulli, binomial, gamma, beta, etc)
- The density function satisfies certain form

$$f(y, \boldsymbol{\theta}) = h(y) \exp\left[\sum_{i=1}^{s} \eta_i(\boldsymbol{\theta}) T_i(y) - b(\boldsymbol{\theta})\right].$$

- After reparameterization (set $\theta_i = \eta_i(\boldsymbol{\theta})$), the canonical form is

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = h(\boldsymbol{y}) \exp\left[\boldsymbol{\theta}^T T(\boldsymbol{y}) - b(\boldsymbol{\theta})\right],$$

where $\boldsymbol{\theta}$ is the (canonical) natural parameter.

Gaussian Example

Consider $x \sim N(\mu, \sigma^2)$.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

which forms a two-parameter exponential family with

$$\theta_1 = \frac{\mu}{\sigma^2},\ \theta_2 = -\frac{1}{2\sigma^2},\ (\theta_1, \theta_2) \in \mathbb{R} \times (-\infty, 0).$$

In this course, we primarily focus on the exponential family with a single canonical parameter. For a random variable $y \sim EF(\theta)$, the density function is

$$f(y, \theta) = h(y) \exp\left(y\theta - b(\theta)\right)$$

where θ is the natural parameter.

Sometimes, we need to consider a more general form with a scale (dispersion) parameter ϕ

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

Gaussian Example

Consider $x \sim N(\mu, \sigma^2)$ where σ^2 is known.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

forms a single-parameter exponential family distribution with the canonical parameter $\theta = \mu$ and the dispersion parameter $\phi = \sigma^2$.

Additional Examples

- Bernoulli(p): $p^x(1-p)^{1-x}$;
- Binomial(n, p);
- Poisson(λ);
- $\chi^2(k)$;
- Gamma(a,b);
- Beta(a,b);
- Negative Binomial(p,m).

Properties of Exponential Family

Consider the exponential family density function,

$$f_Y(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right].$$

- Expectation: $\mathbb{E}(Y) = \mu = b'(\theta)$
- Variance: $\text{var}(Y) = \phi b''(\theta)$
- $b(\theta)$ is always a convex function
- If we express $\text{var}(Y)$ as a function of μ, we get the so-called variance function $V(\mu)$

Example

- Normal distribution $N(\theta, 1)$

- Binary distribution $\mathrm{Bin}(p)$