

# 18.650 – Fundamentals of Statistics

## 7. Generalized linear models

# Linear model

Linear model : -  $Y|X=x \sim \mathcal{N}(\mu(x), \sigma^2 I)$   
-  $\mu(x) = x^\top \beta$  for some unknown  $\beta$ .

GLI:

- $Y|X=x \sim \text{Dist. in exponential family}$ .
- $\mathbb{E}[Y|X=x] = f(x^\top \beta)$ , How to pick  $f$ .

A Gaussian linear model assumes

$$\underbrace{Y|X=x \sim \mathcal{N}}_{\text{1, 假设一个特定族的分布}}(\underbrace{\mu(x)}, \underbrace{\sigma^2 I}),$$

And<sup>1</sup>

$$\mathbb{E}(Y|X=x) = \mu(x) = x^\top \beta,$$

$\uparrow$   
*regression function*

2, 把  $Y|X$  和  $x$  联系起来。（有无限种联系的方式，这里我们只讨论线性的。）

<sup>1</sup>Throughout we drop the boldface notation for vectors

# Components of a linear model

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

1. Random component: the response variable  $Y$  is continuous and  $Y|X = x$  is Gaussian with mean  $\mu(x)$ .
2. Regression function:  $\mu(x) = x^\top \beta$ . linear

# Kyphosis

The Kyphosis data consist of measurements on 81 children following corrective spinal surgery. The binary response variable,  $Y$ , indicates the presence or absence of a postoperative deformity.

The three covariates are:

- ▶  $X^{(1)}$ : Age of the child in month, ↗
- ▶  $X^{(2)}$ : Number of the vertebrae involved in the operation, and ↘
- ▶  $X^{(3)}$ : Start of the range of the vertebrae involved. ↗

Write  $X = (\mathbf{1}, X^{(1)}, X^{(2)}, X^{(3)})^\top \in \mathbb{R}^4$

↑  
mean response ;  
此时X1,X2,X3都是0  
也就是当没有任何信息的时候，  
你的预测值。

# Kyphosis

- The response variable is binary so there is no choice:  
 $Y|X = x$  is Bernoulli with expected value  
 $\mu(x) = \mathbb{E}[Y|X = x] \in (0, 1)$
- We cannot write

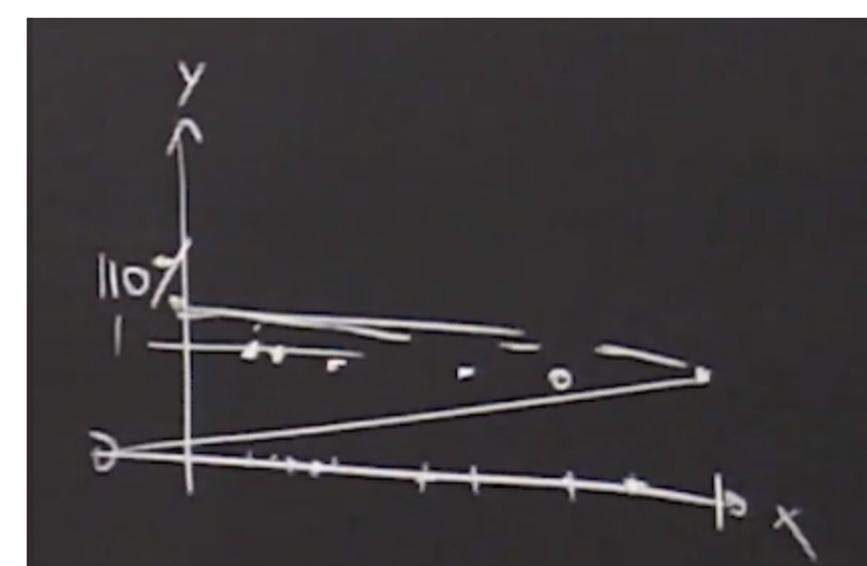
$$\mu(x) = x^\top \beta$$

because the right-hand side ranges through  $\mathbb{R}$

- We need an invertible function  $f$  such that  $f(x^\top \beta) \in (0, 1)$

我们需要的预测值应该符合我们的理解，  
在这个例子中预测值应该在0和1之间。

所以我们需要一个逆函数将我们的Ymap到(0,1)



# Generalization

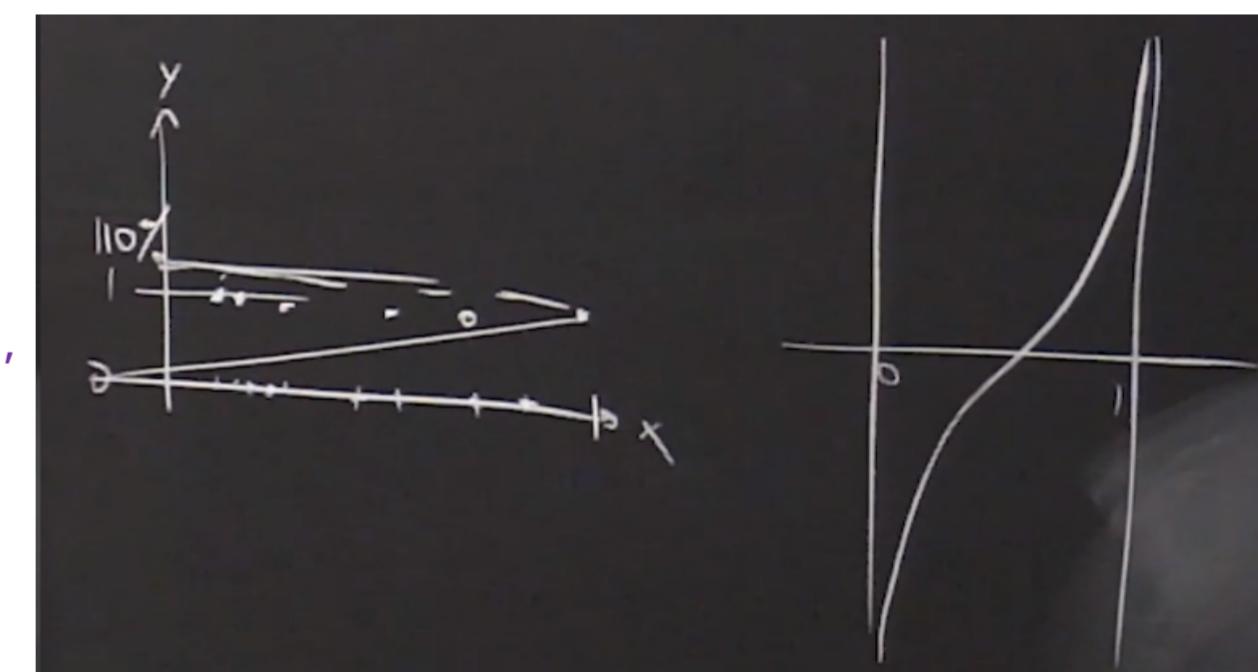
A generalized linear model (GLM) generalizes normal linear regression models in the following directions.

## 1. Random component:

$$Y|X = x \sim \text{some distribution}$$

(e.g. Bernoulli, exponential, Poisson)

左图是一个不符合事实的情况，  
右图将(0,1)map to R



## 2. Regression function:

值域是(0,1)

$$g(\mu(x)) = x^\top \beta$$

其实也就是  $\mu(x) = f(x'\beta)$   
这里  $g$  是  $f$  (上一页) 的反函数

where  $g$  called link function and  $\mu(x) = \mathbb{E}(Y|X = x)$  is the regression function.

# Predator/Prey

Consider the following model for the number of preys  $Y$  that a predator (Hawk) catches per day ~~a predator~~ given a number  $X$  of preys (mice) in its hunting territory.

**Random component:**  $Y > 0$  and the variance of capture rate is known to be approximately equal to its expectation so we propose the following model:

$$Y|X = x \sim \text{Poisss}(\mu(x))$$

very large n, small p

Where  $\mu(x) = \mathbb{E}[Y|X = x]$ .

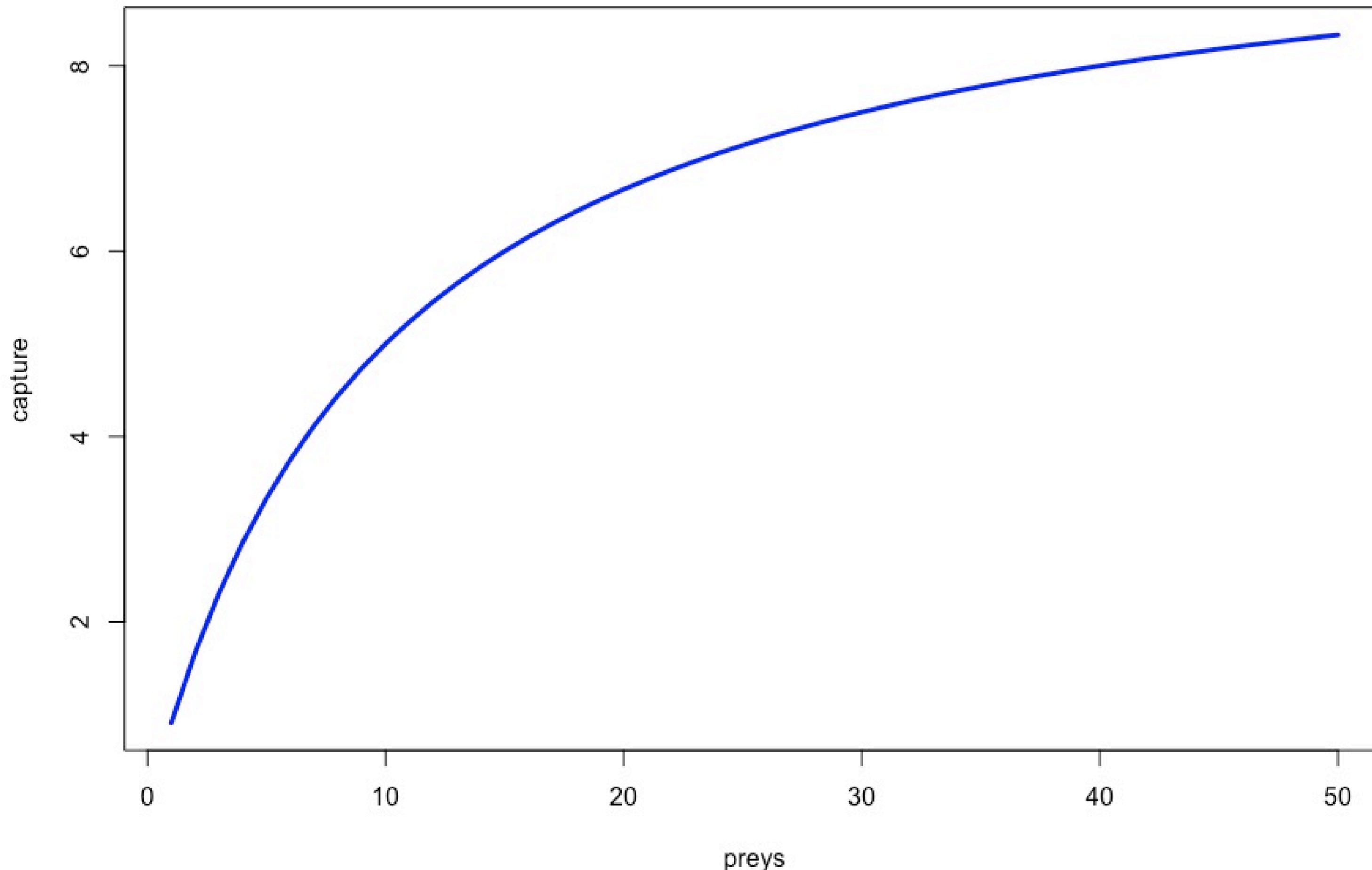
**Regression function:** We assume

$$\mu(x) = \frac{mx}{h + x}, \quad \text{for some unknown } m, h > 0.$$

where:

- ▶  $m$  is the max expected daily preys the predator can cope with
- ▶  $h$  is the number of preys such that  $\mu(h) = \frac{m}{2}$

The regression function  $m(x)$  for  $m = h = 10$



## Example 2: Prey Capture Rate

$$\mu(x) = \frac{m x}{h + x}$$

$$\frac{1}{\mu(x)} = \frac{h + x}{mx}$$

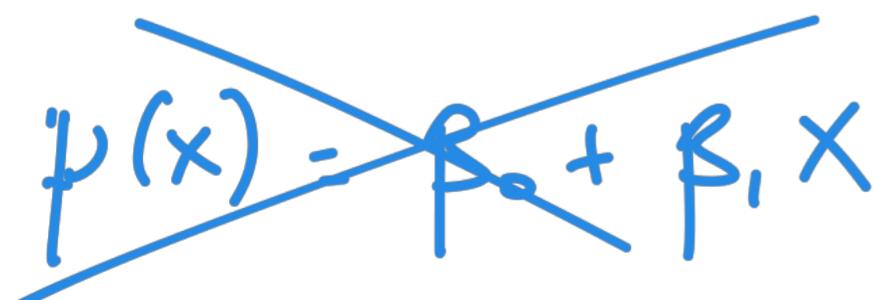
$\mu(x)$  已经确定，此时需要找出一个  $g(x)$ ，  
使得我们最后能得到一个 linear model

不需要将  $[0, \infty)$  maps to  $\mathbb{R}$ , 因为在这个问题中我们知道  $x > 0$ 。

Obviously  $\mu(x)$  is not linear but using **reciprocal link**:  $g(x) = \frac{1}{x}$ ，  
the right-hand side can be made linear in the parameters:

$$g(\mu(x)) = \frac{1}{\mu(x)} = \frac{1}{m} + \frac{h}{m} \cdot \frac{1}{x} = \beta_0 + \beta_1 \frac{1}{x}.$$

这种形式的回归方程不行是因为和问题不兼容



# Exponential Family

A family of distribution  $\{\mathbb{P}_\theta : \theta \in \Theta\}$ ,  $\Theta \subset \mathbb{R}^k$  is said to be a  **$k$ -parameter exponential family** on  $\mathbb{R}^q$ , if there exist real valued functions:

- ▶  $\eta_1, \eta_2, \dots, \eta_k$  and  $B$  of  $\theta$ ,  
我们通过regression function和X的数据来构造这个B( $\theta$ )
- ▶  $T_1, T_2, \dots, T_k$ , and  $h$  of  $y \in \mathbb{R}^q$  such that the density function (pmf or pdf) of  $\mathbb{P}_\theta$  can be written as

$$f_\theta(y) = \exp \left[ \sum_{i=1}^k \eta_i(\theta) T_i(y) - B(\theta) \right] h(y)$$

*k : theta的维度*

*( $\eta_1(\theta), \dots, \eta_k(\theta)$ )*  $\begin{pmatrix} T_1(y) \\ \vdots \\ T_k(y) \end{pmatrix}$  normalized term

*$f_\theta(y) = \exp(\eta(\theta)^T T(y) - B(\theta)) h(y)$*

*$f_\theta(y) = \exp(\eta(\theta)^T T(y)) e^{-B(\theta)} h(y)$*  形式有多种写法

## Normal distribution example

- Consider  $Y \sim \mathcal{N}(\mu, \sigma^2)$ ,  $\theta = (\mu, \sigma^2)$ . The density is

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

which forms a two-parameter exponential family with

可以任意分解，但是不要这样做

$$\eta_1 = \frac{\mu}{\sigma^2}, \quad \eta_2 = -\frac{1}{2\sigma^2}, \quad T_1(y) = y, \quad T_2(y) = y^2,$$

$$B(\theta) = \frac{\mu^2}{2\sigma^2} + \log(\sigma\sqrt{2\pi}), \quad h(y) = 1.$$

- When  $\sigma^2$  is known, it becomes a one-parameter exponential family on  $\mathbb{R}$ :

$$\eta = \frac{\mu}{\sigma^2}, \quad T(y) = y, \quad B(\theta) = \frac{\mu^2}{2\sigma^2}, \quad h(y) = \frac{e^{-\frac{y^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}.$$

# Examples of discrete distributions

The following distributions form **discrete** exponential families of distributions with **pmf**

- Bernoulli( $p$ ):  $p^y(1-p)^{1-y}$ ,  $y \in \{0, 1\}$

$$\begin{aligned} & \exp(y \lg p + (1-y) \lg(1-p)) \\ &= \exp\left(\underbrace{y \lg\left(\frac{P}{1-P}\right)}_{T(y)} - \underbrace{(-g(1-p))}_{B(p)}\right) \cdot \underbrace{\frac{1}{h(y)}}_{h(y)} \end{aligned}$$

- Poisson( $\lambda$ ):  $\frac{\lambda^y}{y!} e^{-\lambda}$ ,  $y = 0, 1, \dots$

$$\text{Poisson } \frac{\lambda^y}{y!} e^{-\lambda} = \exp\left(\underbrace{-\lambda}_{B(\lambda)} + \underbrace{y \log \lambda}_{T(y)}\right) \frac{1}{y!}$$

# Examples of Continuous distributions

The following distributions form **continuous** exponential families of distributions with **pdf**:

- ▶ Gamma( $a, b$ ):  $\frac{1}{\Gamma(a)b^a}y^{a-1}e^{-\frac{y}{b}}$ ;
  - ▶ above:  $a$ : shape parameter,  $b$ : scale parameter
  - ▶ reparametrize:  $\mu = ab$ : mean parameter

$$\frac{1}{\Gamma(a)} \left(\frac{a}{\mu}\right)^a y^{a-1} e^{-\frac{ay}{\mu}}.$$

- ▶ Inverse Gamma( $\alpha, \beta$ ):  $\frac{\beta^\alpha}{\Gamma(\alpha)}y^{-\alpha-1}e^{-\beta/y}$ .
- ▶ Inverse Gaussian( $\mu, \sigma^2$ ):  $\sqrt{\frac{\sigma^2}{2\pi y^3}} e^{\frac{-\sigma^2(y-\mu)^2}{2\mu^2 y}}$ .

Others: Chi-square, Beta, Binomial, Negative binomial distributions.

# One-parameter canonical exponential family

- Canonical exponential family for  $k = 1$ ,  $y \in \mathbb{R}$

减号

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

for some *known* functions  $b(\cdot)$  and  $c(\cdot, \cdot)$ .

- If  $\phi$  is known, this is a one-parameter exponential family with  $\theta$  being the canonical parameter.
- If  $\phi$  is unknown, this may/may not be a two-parameter exponential family.
- $\phi$  is called **dispersion parameter**.
- In this class, we always assume that  $\phi$  is *known*.

$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi}\right) \cdot h(y) \quad \text{where } h(y) = e^{c(y, \phi)}$$

## Normal distribution example

- ▶ Consider the following Normal density function with known variance  $\sigma^2$ ,

$$\begin{aligned}f_{\theta}(y) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\&= \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\},\end{aligned}$$

- ▶ Therefore  $\theta = \underline{\mu}$ ,  $\phi = \underline{\sigma^2}$ ,  $b(\theta) = \underline{\frac{\theta^2}{2}}$ , and

$$c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right).$$

# Other distributions

Table 1: Exponential Family

	Normal	Poisson	Bernoulli
Notation	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{P}(\mu)$	$\mathcal{B}(p)$
Range of $y$	$(-\infty, \infty)$	$[0, -\infty)$	$\{0, 1\}$
$\phi$	$\sigma^2$	1	1
$b(\theta)$	$\frac{\theta^2}{2}$	$e^\theta$	$\log(1 + e^\theta)$
$c(y, \phi)$	$-\frac{1}{2} \left( \frac{y^2}{\phi} + \log(2\pi\phi) \right)$	$-\log y!$	0

# Likelihood

Let  $\ell(\theta) = \log f_\theta(Y)$  denote the log-likelihood function.

The mean  $\mathbb{E}(Y)$  and the variance  $\text{var}(Y)$  can be derived from the following identities

► First identity

$$\mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right) = \textcircled{o}$$

$$\begin{aligned} \mathbb{E}\left[\frac{d\ell}{d\theta}\right] &= \int_Y \frac{d\ell}{d\theta} f_\theta(y) dy \\ &\xrightarrow{\text{微分和导数互换 (在这里可以)}} = \int_Y \frac{d \ln f_\theta(y)}{d\theta} f_\theta(y) dy \\ &= \int_Y \frac{1}{f_\theta(y)} \frac{df_\theta(y)}{d\theta} f_\theta(y) dy \\ &= \frac{d}{d\theta} \int_Y f_\theta(y) dy = \frac{d}{d\theta} 1 = 0. \end{aligned}$$

► Second identity

$$\mathbb{E}[\ell'(\theta)] = \mathbb{E}\left[\frac{\frac{\partial}{\partial \theta} f_\theta(X)}{f_\theta(X)}\right] = \int_{-\infty}^{\infty} \left(\frac{\frac{\partial}{\partial \theta} f_\theta(x)}{f_\theta(x)}\right) f_\theta(x) dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f_\theta(x) dx = 0$$

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) + \mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right)^2 = 0. \quad \text{Var}(\ell'(\theta)) = \mathbb{E}[\ell'(\theta)^2] - \mathbb{E}[\ell'(\theta)]^2 \\ = \mathbb{E}[(\ell'(\theta))^2]$$

$$\Leftrightarrow \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)'\right] = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] \quad (\text{minima})$$

f.i.      F.I.

两种fisher information的表达方式

$$\begin{aligned} &= \mathbb{E}\left[\left(\frac{\frac{\partial}{\partial \theta} f_\theta(X)}{f_\theta(X)}\right)^2\right] \\ &= \int_{-\infty}^{\infty} \frac{\left(\frac{\partial}{\partial \theta} f_\theta(x)\right)^2}{f_\theta(x)} dx \end{aligned}$$

## Expected value

Note that

$$\ell(\theta) = \frac{Y\theta - b(\theta)}{\phi} + c(Y; \phi),$$

Therefore

$$\frac{\partial \ell}{\partial \theta} = \frac{Y - b'(\theta)}{\phi}$$

It yields

$$0 = \mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right) = \frac{\mathbb{E}(Y) - b'(\theta)}{\phi},$$

which leads to

$$\mathbb{E}(Y) = b'(\theta)$$

# Variance

On the other hand we have we have

$$\frac{\partial^2 \ell}{\partial \theta^2} + \left( \frac{\partial \ell}{\partial \theta} \right)^2 = -\frac{b''(\theta)}{\phi} + \left( \frac{Y - b'(\theta)}{\phi} \right)^2$$

and from the previous result,

$$\frac{Y - b'(\theta)}{\phi} = \frac{Y - \mathbb{E}(Y)}{\phi}$$

Together, with the second identity, this yields

$$0 = -\frac{b''(\theta)}{\phi} + \frac{\text{var}(Y)}{\phi^2},$$

which leads to

$$\text{var}(Y) = b''(\theta) \cdot \phi$$

## Example: Poisson distribution

$$b(\theta) = p = e^\theta$$

$$\log p = \theta \Rightarrow p = e^\theta$$

Example: Consider a Poisson likelihood,

这里的mu是Y的参数

$$f(y) = \frac{\mu^y}{y!} e^{-\mu} = \exp \left( y \underbrace{\log \mu}_{\theta} - \underbrace{\mu - \log(y!)}_{c(y, \phi)} \right)$$

Thus,

mu不是回归方程的参数，log(mu)才是  
这样做转化成一个GLM

$$\theta = \log \mu \quad b(\theta) = e^\theta \quad \phi = 1 \quad c(y, \phi) = -\log(y!),$$

So

$$\mu = e^\theta, \quad b(\theta) = p \quad b''(\theta) = p$$

$$\mathbb{E}[X] = p, \quad \text{Var}[X] = p \cdot 1 = p$$

# Link function

Linear model: -  $Y|X=x \sim N(\mu(x), \sigma^2 I)$   
-  $\mu(x) = x^\top \beta$  for some unknown  $\beta$ .

GLM:

- $Y|X=x \sim \text{Dist. in exponential fam.}$
- $\mathbb{E}[Y|X=x] = f(x^\top \beta)$ , How to pick  $f$ ?

- $\beta$  is the parameter of interest, and needs to appear somehow in the likelihood function to use maximum likelihood.
- A link function  $g$  relates the linear predictor  $X^\top \beta$  to the mean parameter  $\mu$ ,

mu在上一页中，是Y的参数，我们要用 $X'\beta$ 来表达mu  
用predictor表达parameter

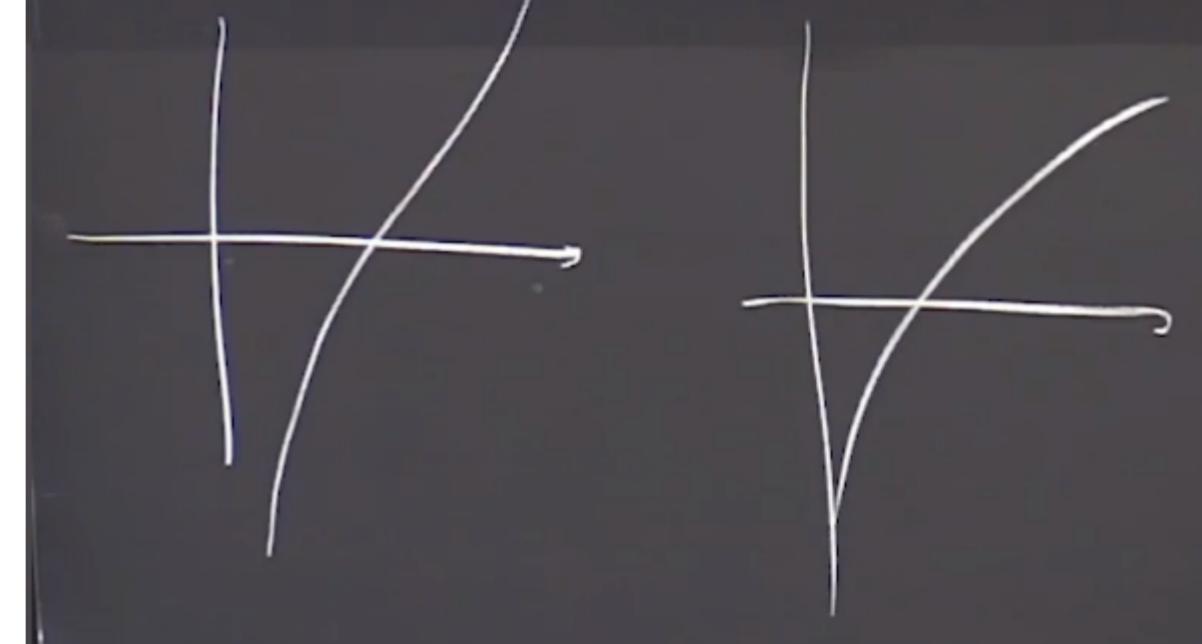
$$X^\top \beta = g(\mu). = g(\mu(X))$$

- $g$  is required to be monotone increasing and differentiable

invertible

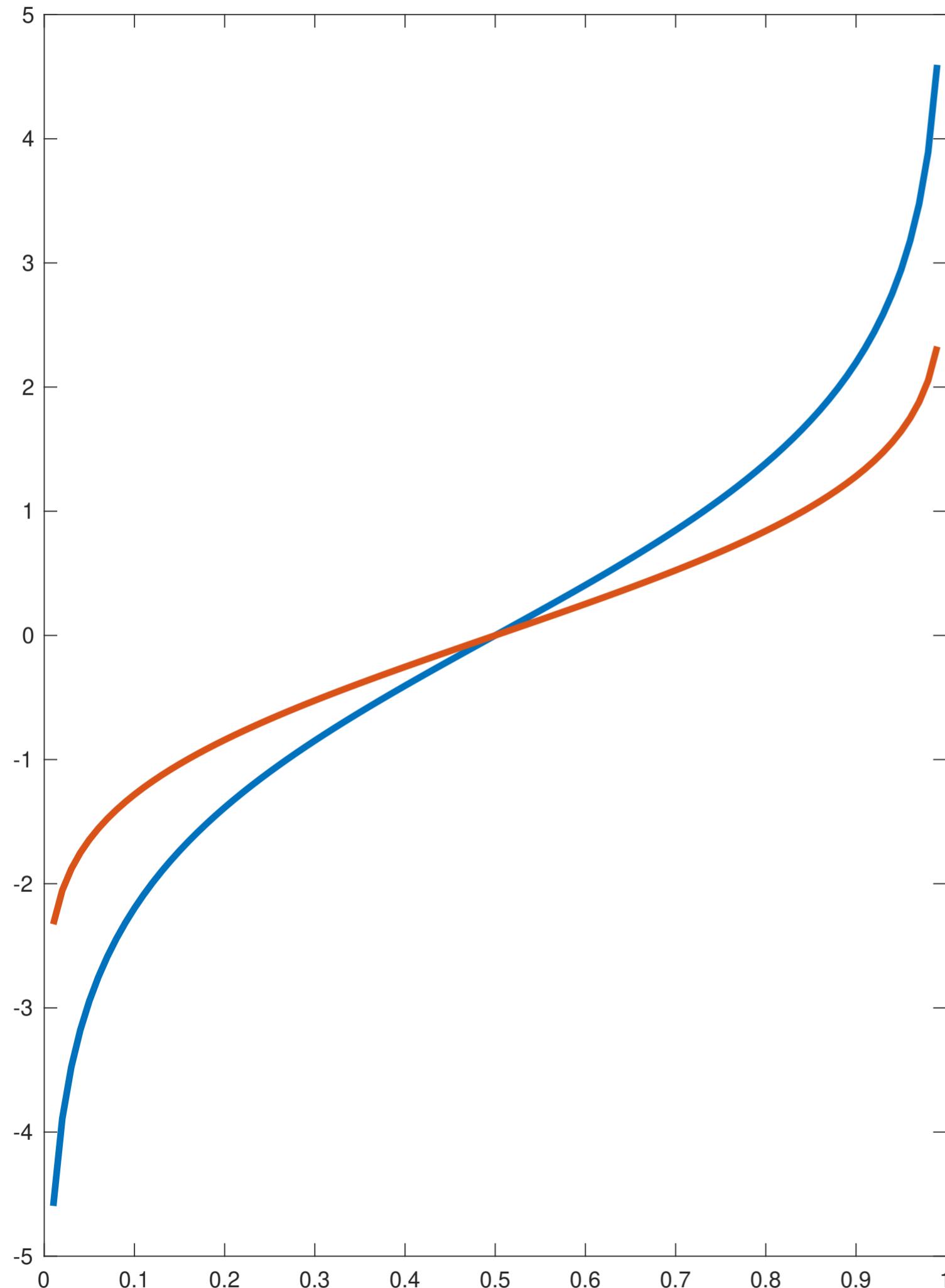
$$\mu = g^{-1}(X^\top \beta).$$

# Examples of link functions



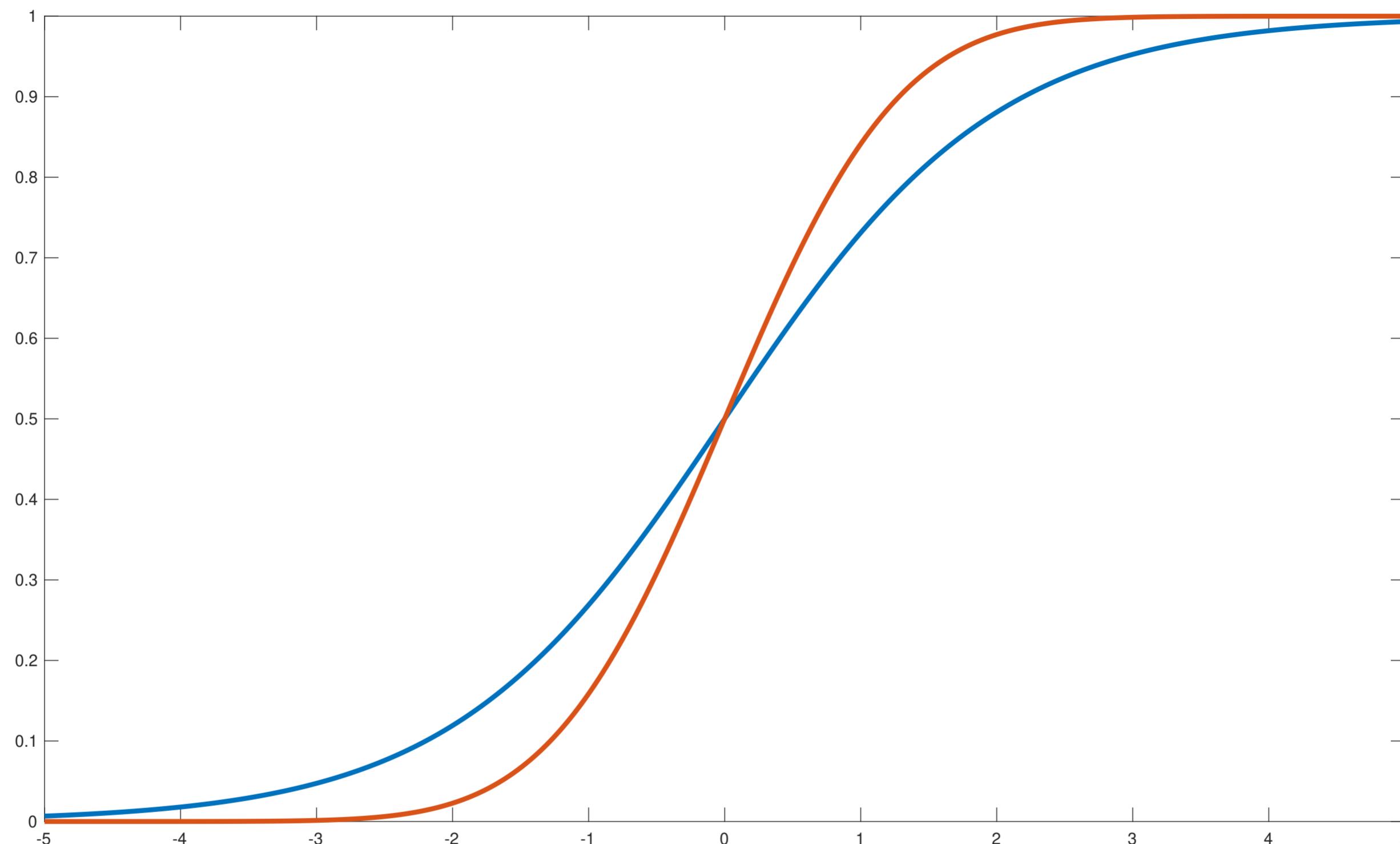
- For LM,  $g(\cdot) = \text{identity}$ . ✓
- Poisson data. Suppose  $Y|X \sim \text{Poisson}(\mu(X))$ .
  - $\mu(X) > 0$ ; 把(0,inf)的数据map到R
  - $\log(\mu(X)) = X^\top \beta$ ; log-link
  - In general, a link function for the count data should map  $(0, +\infty)$  to  $\mathbb{R}$ .
  - The log link is a natural one.
- Bernoulli/Binomial data.
  - $0 < \mu < 1$ ;
  - $g$  should map  $(0, 1)$  to  $\mathbb{R}$ :
  - 3 choices:
    1. logit:  $\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = X^\top \beta$ ; 先把(0,1)map到 $\mathbb{R}_+$ , 再用log map到R
    2. probit:  $\Phi^{-1}(\mu(X)) = X^\top \beta$  where  $\Phi(\cdot)$  is the normal cdf;
  - The logit link is the natural choice.

# Examples of link functions for Bernoulli response



- in blue:  
$$g_1(x) = f_1^{-1}(x) = \log\left(\frac{x}{1-x}\right)$$
 (logit link)
- in red:  
$$g_2(x) = f_2^{-1}(x) = \Phi^{-1}(x)$$
 (probit link)

# Examples of link functions for Bernoulli response



- in blue:  $f_1(x) = \frac{e^x}{1 + e^x}$
- in red:  $f_2(x) = \Phi(x)$  (Gaussian CDF)

# Canonical Link

- The function  $g$  that links the mean  $\mu$  to the canonical parameter  $\theta$  is called **Canonical Link**:

$$g(\mu) = \theta$$

- Since  $\mu = b'(\theta)$ , the canonical link is given by

b是构造出来的函数  
g是link函数

$$b'(g(\mu)) = \mu$$

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

- If  $\phi > 0$ , the canonical link function is **strictly increasing**.  
Why?

$$g \text{ str. } \Leftrightarrow g^{-1} \text{ str. } \Leftrightarrow (g^{-1})' > 0 \Leftrightarrow b'' > 0 \Leftrightarrow \text{Var} > 0$$

$\phi > 0$

## Example: the Bernoulli distribution

- We can check that

$$b(\theta) = \log(1 + e^\theta)$$

- Hence we solve

$$b'(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} = \mu \quad \Leftrightarrow \quad \theta = \log\left(\frac{\mu}{1-\mu}\right)$$

- The canonical link for the Bernoulli distribution is the *logit link*

# Other examples

$$\mu = b'(\theta)$$

link functions

前三个mean都正好等于参数

	$b(\theta)$	$g(\mu)$
Normal	$\theta^2/2$	$\mu$
Poisson	$\exp(\theta)$	$\log \mu$
Bernoulli	$\log(1 + e^\theta)$	$\log \frac{\mu}{1-\mu}$
Gamma	$-\log(-\theta)$	$-\frac{1}{\mu}$

exercise

# Model and notation

$X_i$ 是一个p维的观测变量,  $Y_i$ 是一个预测变量  
通过 $X_i$ 预测 $Y_i$ 一共有n对 $X_i$ 和 $Y_i$

- Let  $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ ,  $i = 1, \dots, n$  be independent random pairs such that the conditional distribution of  $Y_i$  given  $X_i = x_i$  has density in the canonical exponential family:

给了Y的分布， $b$ , phi  
我们就知道了任何 $Y|X$ 的信息

$$f_{\theta_i}(y_i) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}.$$

每个 $Y_i$ 有不同的模型， $\theta$ 将所有的 $Y_i$ 和 $X_i$ 联系起来。  
 $\theta_i = X_i^\top \beta$

$\beta$ 是唯一的未知参数，当我们有Y和X的数据的后，用这些数据来估计 $\beta$

- $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ ,  $\mathbf{X} = (X_1, \dots, X_n)^\top = \begin{pmatrix} \vdots & \vdots \\ X_1 & \dots \\ \vdots & \vdots \end{pmatrix} \in \mathbb{R}^{n \times p}$
- Here the mean  $\mu_i = \mathbb{E}[Y_i | X_i]$  is related to the canonical parameter  $\theta_i$  via

$$\mu_i = b'(\theta_i)$$

这里的 $\theta$ 是canonical模型的 $\theta$ ，我们通过构建canonical模型，将 $\mu$ 转换成 $\theta$   
那么，我们用反函数，也能把 $\mu$ 找回来。  
这么做的原因也是指数分布族的优点

- and  $\mu_i$  depends linearly on the covariates through a link function  $g$ :

$$g(\mu_i) = X_i^\top \beta.$$

我们要将Y的期望和 $\beta X$ 通过link function联系起来。

$$b'(g(\mu)) = \mu$$

## Back to $\beta$

likelihood function里面只有 $\theta$   
我们想要的是 $\beta$ ，所以我们把 $\theta$ 替换成 $\beta$

- Given a link function  $g$ , note the following relationship between  $\beta$  and  $\theta$ :

$$\begin{aligned} g(\mu) &= \theta \\ \mu &= b'(\theta) \end{aligned} \quad \begin{aligned} \theta_i &= (b')^{-1}(\mu_i) \\ &= (b')^{-1}(g^{-1}(X_i^\top \beta)) \equiv \underline{h(X_i^\top \beta)}, \end{aligned}$$

where  $h$  is defined as

$$h = (b')^{-1} \circ g^{-1} = (g \circ b')^{-1}.$$

- Remark: if  $g$  is the **canonical** link function,  $h$  is *the identity*

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

# Log-likelihood

- The log-likelihood is given by

$$\begin{aligned}\ell_n(\mathbf{Y}, \mathbb{X}, \beta) &= \sum_i \frac{Y_i \theta_i - b(\theta_i)}{\phi} + cst \\ &\stackrel{\theta_i = \Theta_i(X_i) = h(X_i^\top \beta)}{=} \sum_i \frac{Y_i h(X_i^\top \beta) - b(h(X_i^\top \beta))}{\phi} + cst\end{aligned}$$

up to a constant term.

- Note that when we use the canonical link function, we obtain the simpler expression

$$\ell_n(\mathbf{Y}, \mathbb{X}, \beta) = \sum_i \frac{Y_i X_i^\top \beta - b(X_i^\top \beta)}{\phi} + cst$$

$\square H_\beta \ell_n = \sum_i H_\beta \frac{b(X_i^\top \beta)}{\phi} = \frac{1}{\phi} \sum_{i=1}^n b''(X_i^\top \beta) X_i X_i^\top \left[ -\frac{X_i^\top H_\beta}{\phi} X_i \right] = \frac{1}{\phi} \sum_{i=1}^n b''(X_i^\top \beta) (X_i^\top X_i)$

# Strict concavity

- ▶ The log-likelihood  $\ell(\theta)$  is strictly concave using the canonical function when  $\phi > 0$ . Why? if Rank(  $\mathbf{X}$  ) = p  
full rank
- ▶ As a consequence the maximum likelihood estimator is unique
- ▶ On the other hand, if another parameterization is used, the likelihood function may not be strictly concave leading to **several local maxima**.

$$\nabla_{\beta} b(X_i^T \beta) = b'(X_i^T \beta) X_i$$

$$H_{\beta} b(X_i^T \beta) = b''(X_i^T \beta) X_i X_i^T$$

$$\frac{1}{N} \sum_{i=1}^N Y_i X_i^T \beta - b(X_i^T \beta)$$

$$\nabla_{\beta} = 0 \Leftrightarrow \sum_{i=1}^N Y_i \boxed{X_i} - b'(X_i^T \beta) \boxed{X_i} = 0$$

solve the p equations  
if  $b'$  is Gaussian canonical, it's LSE.  
if  $b'$  is non-linear, it's nasty  
so we use optimization algorithms.

# Concluding remarks

- ▶ Maximum likelihood for Bernoulli  $Y$  and the logit link is called *logistic regression*.
- ▶ In general, there is **no closed form** for the MLE and we have to use *optimization algorithms* (*gradient ascent*)  
don't have Gaussian error, so rely on CLT
- ▶ The asymptotic normality of the MLE also applies to GLMs.

测试 $\beta$ 是不是显著不等于0