18.650 – Fundamentals of Statistics

7. Generalized linear models

Linear model

A Gaussian linear model assumes

$$Y|X = x \sim \mathcal{N}(\mu(x), \sigma^2 I),$$

And[1]

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

[1] 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: $\mu(x) = x^\top \beta$.

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 deforming.

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

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 $f(x^\top \beta) \in (0, 1)$

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)

2. Regression function:

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

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:

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

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

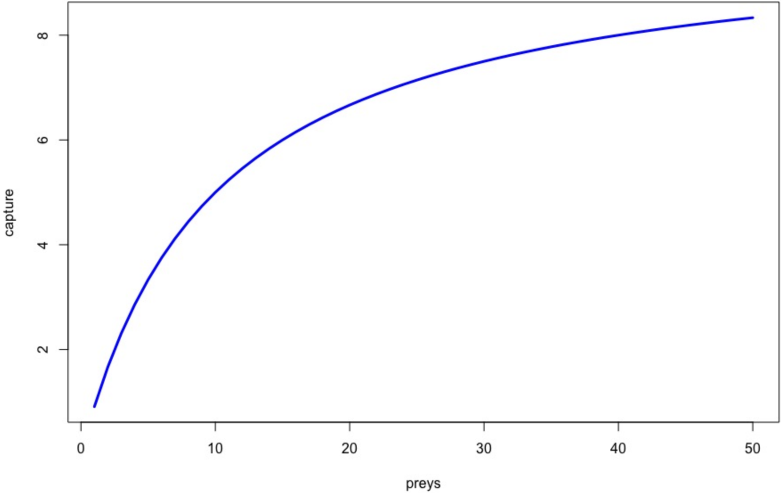
Regression function: We assume

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

where:

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

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



Example 2: Prey Capture Rate

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

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

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:

- $\eta_1, \eta_2, \cdots, \eta_k$ and B of θ,
- T_1, $T_2, \cdots, T_k$, and h of $y \in \mathbb{R}^q$ such that the density function (pmf or pdf) of $\mathbb{P}_\theta$ can be written as

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

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

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

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

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

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

Examples of discrete distributions

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

- Bernoulli(p): $\quad p^y(1-p)^{1-y}$, $y \in \{0,1\}$
- Poisson(λ): $\dfrac{\lambda^y}{y!}e^{-\lambda}$, $y = 0, 1, \ldots$.

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(α, β): $\frac{\beta^\alpha}{\Gamma(\alpha)} y^{-\alpha-1} e^{-\beta/y}$.
- Inverse Gaussian(μ, σ^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 $k = 1$, $y \in \mathbb{R}$

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

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

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

Normal distribution example

- Consider the following Normal density function with known variance σ^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 $\theta = \mu,\ \phi = \sigma^2,\ b(\theta) = \frac{\theta^2}{2}$, and

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

Other distributions

Table 1: Exponential Family

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

Likelihood

Let $\ell(\theta) = \log f_\theta(Y)$ denote the log-likelihood function.
The mean $\mathbb{E}(Y)$ and the variance $\mathrm{var}(Y)$ can be derived from the following identities

- First identity
$$\mathbb{E}(\frac{\partial \ell}{\partial \theta}) = 0$$
- Second identity
$$\mathbb{E}(\frac{\partial^2 \ell}{\partial \theta^2}) + \mathbb{E}(\frac{\partial \ell}{\partial \theta})^2 = 0.$$

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

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

which leads to

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

Variance

On the other hand we have we have

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

and from the previous result,

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

Together, with the second identity, this yields

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

which leads to

$$\mathsf{var}(Y) = b''(\theta)\phi.$$

Example: Poisson distribution

Example: Consider a Poisson likelihood,

$$f(y) = \frac{\mu^y}{y!} e^{-\mu} = \exp\big(y \log \mu - \mu - \log(y!)\big)$$

Thus,

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

So

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

Link function

- β 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 μ,

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

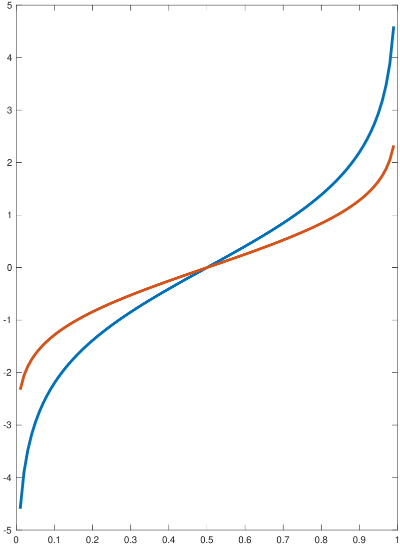
- g is required to be monotone increasing and differentiable

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

Examples of link functions

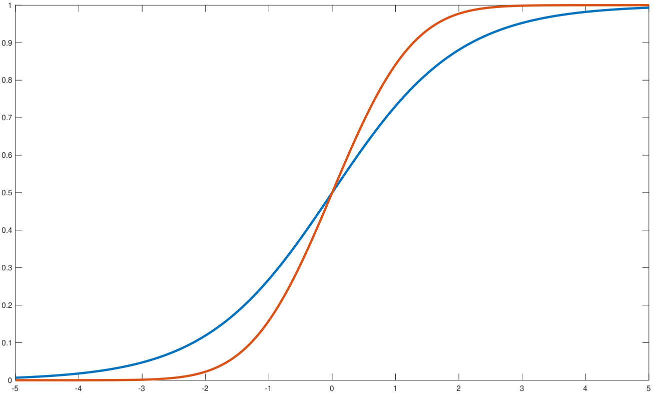
- For LM, $g(\cdot) =$ identity.
- Poisson data. Suppose $Y|X \sim \text{Poisson}(\mu(X))$.
  - $\mu(X) > 0$;
  - $\log(\mu(X)) = X^\top \beta$;
  - In general, a link function for the count data should map $(0, +\infty)$ to $\mathbb{R}$.
  - The log link is a natural one.
- Bernoulli/Binomial data.
  - $0 < \mu < 1$;
  - g should map $(0, 1)$ to $\mathbb{R}$:
  - 3 choices:
    1. logit: $\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = X^\top \beta$;
    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: $f_1(x) = \dfrac{e^x}{1+e^x}$
- in red: $f_2(x) = \Phi(x)$ (Gaussian CDF)

Canonical Link

- The function g that links the mean μ to the canonical parameter θ is called **Canonical Link**:

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

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

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

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

Example: the Bernoulli distribution

- We can check that

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

- Hence we solve

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

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

Other examples

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

Model and notation

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

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

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, $\mathbb{X} = (X_1, \ldots, X_n)^\top$
- Here the mean $\mu_i = \mathbb{E}[Y_i|X_i]$ is related to the canonical parameter θ_i via

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

- and μ_i depends linearly on the covariates through a link function g:

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

Back to β

- Given a link function g, note the following relationship between β and θ:

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

  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 **identity**.

Log-likelihood

- The log-likelihood is given by

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

  up to a constant term.

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

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

Strict concavity

- The log-likelihood $\ell(\theta)$ is **strictly concave** using the canonical function when $\phi > 0$. `Why?`
- 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**.

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
- The asymptotic normality of the MLE also applies to GLMs.