Generalized Linear Models

Jiajun Zhang*

January 31, 2026

Contents

*Department of Mathematics and Statistics, jiajun.zhang@mail.mcgill.ca, McGill ID: 261105766

1 Introduction

Suppose we have the following experiment:

- We have data on 79 cervical cancer patients, and we want to study the association between histology (low v.s high) and the risk factors (neovasculation, pulsatility, endometrium height).
- We would like to use regression methods by coding histology as 0 (low) and 1 (high) and using those 3 factors as covariates.

There are many potential issues with this experiment. OLS is genius, but potentially inefficient or awkward to work with. One alternative is to first observe that the response variable Y_i is Bernoulli (0 or 1), so we first assume

$$\mathbb{P}(Y_i = 1|\boldsymbol{x}_i) = p(x_1, x_2, x_3|\beta),$$

where $p_i(\beta) \equiv p(x_1, x_2, x_3|\beta)$ is any function of the three covariates and an unknown parameter vector β. The likelihood of the n observations is

$$L(y_1, \cdots, y_n; \beta) = \prod_{i=1}^{n} p_i(\beta)^{y_i} \cdot (1 - p_i(\beta))^{1-y_i},$$

and the MLE is given by

$$\widehat{\beta} = \arg\max_{\beta} \prod_{i=1}^{n} p_i(\beta)^{y_i} \cdot (1 - p(y_i))^{1-y_i}.$$

For a particular choice of p, we can then try to maximize this either analytically or numerically. The first choice for p is the CDF of a $\mathcal{N}(0, 1)$ random variable:

$$p_i = \Phi(\widehat{z}) = \int_{-\infty}^{\widehat{z}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du,$$

and $\Phi^{-1}(p_i)$ is often referred to as the probit function. The standard approach is to then assume

$$\Phi^{-1}(p_i) = \beta_0 + \beta_1 x_i.$$

The disadvantages for this model is that is it computationally challenging, also not an obvious interpretation of parameters for even a univariate model.

An alternative approach is to use the *logit*:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

Under this model, we have

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x.$$

2 Logistic Regression

2.1 Understanding logistic regression

In the logistic regression model,

$$p_i = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

Under this model, we have

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x.$$

We call $\frac{p_i}{1-p_i}$ the *odds* of $Y_i = 1$. β_0 is the *log-odds* for an observation with $x = 0, \exp(\beta_0)$ would be the odds ratio for such an observation. β_1 is thus the increase in the log-odds of a positive response for a one unit increase in x, $\exp(\beta_1)$ is the *relative* increase in odds.

We can center the covariate x by using $x_c = x - \bar{x}$ which yields

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0^* + (x_i - \bar{x})\beta_1, \quad \beta_0^* = \beta + \bar{x}\beta_1.$$

Take two observations, indexed i, j with covariate x and $x + 1$, then the risk ratio is

$$\begin{aligned}
\frac{\mathbb{P}(X_j = 1)}{\mathbb{P}(X_i = 1)} &= \frac{p_j}{p_i} \\
&= \frac{\exp(\beta_0 + (x+1)\beta_1)}{1 + \exp(\beta_0 + (x+1)\beta_1)} \cdot \frac{1 + \exp(\beta_0 + x\beta_1}{\exp(\beta_0 + x\beta_1)} \\
&= \exp(\beta_1) \cdot \frac{1 + \exp(\beta_0 + x\beta_1)}{1 + \exp(\beta_0 + (x+1)\beta_1)} \\
&\approx \exp(\beta_1),
\end{aligned}$$

when p_i, p_j are small. Therefore, for small values of p_i (when the response probability is small), then the odds ratio $\exp(\beta_1)$ can be interpreted as an approximate relative increase in the probability of response for a one unit increase in x. In many applications, this is true.

2.2 Asymptotic Normality of the MLE

In a univariate logistic regression model, recall that

$$p_i = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

The response Y_i is binary: $0, 1$, so the likelihood of the sample is given by

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i(\beta_0, \beta_1, x)^{y_i} \cdot (1 - p_i(\beta_0, \beta_1, x))^{1-y_i},$$

with the log-likelihood given by

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i \log p(\beta_0, \beta_1, x) + (1 - y_i) \log(1 - p(\beta_0, \beta_1, x)). \tag{2.1}$$

Suppose we can maximize the log-likelihood defined in (2.1) and get the maximum likelihood estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\beta_0}, \widehat{\beta_1})$:

$$(\widehat{\beta_0}, \widehat{\beta_1}) := \arg\max_{\beta_0, \beta_1} \sum_{i=1}^{n} y_i \log p(\beta_0, \beta_1, x) + (1 - y_i) \log(1 - p(\beta_0, \beta_1, x)).$$

The MLE estimate $\widehat{\boldsymbol{\theta}} = (\widehat{\beta_0}, \widehat{\beta_1})$ can then be the estimated coefficients for univariate logistic regression. In asymptotic theory, the maximum likelihood estimator is asymptotically normal. We first introduce the regularity conditions:

Definition 1 (Regularity Conditions)**.** *In a random sample $X_1, \cdots, X_n \sim f(x, \theta)$ where θ is one-dimensional, we give the following regularity conditions:*

(i) The family $\left\{ f(x, \theta) : \theta \in \Theta \subset \mathbb{R} \right\}$ has a common support that does not depend on θ.

(ii) $\dfrac{d}{d\theta} \log f(x, \theta)$ always exists.

(iii) For any statistic $h(\boldsymbol{X}) \in L^1$,

$$\frac{d}{d\theta} \int_{\mathcal{S}} h(\boldsymbol{x}) f(\boldsymbol{x}, \theta) dx = \int_{\mathcal{S}} h(\boldsymbol{x}) \frac{d}{d\theta} f(\boldsymbol{x}, \theta) dx$$

Note that the likelihood cannot (except in certain cases) be maximized analytically, so one must resort to numerical methods. But we will talk about that later. Assume for now that we can obtain ML estimate $\widehat{\boldsymbol{\theta}}$ somehow, and let the true distribution that generates the data be $q(y)$.

Theorem 1. *If there exists a parameter vector $\boldsymbol{\theta}_0$ such that $p_1(\boldsymbol{\theta}_0) = q(y_i)$ for all observations, where q is the true distribution. Then the MLE of $\boldsymbol{\theta}$ converges to $\boldsymbol{\theta}_0$ in probability, and moreover*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathcal{J}^{-1}(\boldsymbol{\theta}_0)\right),$$

where

$$\Big(\mathcal{J}(\boldsymbol{\theta}_0)\Big)_{ij} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta}, y)}{\partial\theta_i \, \partial\theta_j}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

We will prove the theorem in a univariate setting. By i.i.d assumption, we can write the log-likelihood as

$$\ell_n(\theta, y) = \sum_{i=1}^{n} \ell(\theta, y_i).$$

Let $\dot{l}_n(\theta, y)$ be the first derivative of the log-likelihood with respect to θ for n i.i.d observations. We then do a Taylor expansion for $\ell_n(\dot{\theta}, y)$ around θ_0:

$$\dot{\ell}_n(\theta, y) = \dot{\ell}_n(\theta_0, y) + (\theta - \theta_0) \cdot \ddot{\ell}_n(\theta_0, y) + \frac{(\theta - \theta_0)^2}{2} \cdot \dddot{\ell}_n(\theta^*, y), \quad \theta^* \text{ is between } \theta, \theta_0.$$

Since $\widehat{\theta}_n$ is the MLE, so $\dot{\ell}_n(\widehat{\theta}_n, y) = 0$, hence

$$0 = \dot{\ell}_n(\theta_0, y) + (\widehat{\theta}_n - \theta_0) \cdot \ddot{\ell}_n(\theta_0, y) + \frac{(\widehat{\theta}_n - \theta_0)^2}{2} \cdot \dddot{\ell}_n(\theta^*, y),$$

which implies

$$-\dot{\ell}_n(\theta_0, y) = (\widehat{\theta}_n - \theta_0) \cdot \ddot{\ell}_n(\theta_0, y) + \frac{(\widehat{\theta}_n - \theta_0)^2}{2} \cdot \dddot{\ell}_n(\theta^*, y),$$

hence

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{\dot{\ell}_n(\theta_0, y)/\sqrt{n}}{-\frac{1}{n}\ddot{l}_n(\theta_0, y) - \frac{(\widehat{\theta}_n - \theta_0)}{2n} \cdot \dddot{\ell}_n(\theta^*, y)}. \tag{2.2}$$

To understand better, we need to know the following properties of $\dot{\ell}(\theta, y_i)$:

- $\mathbb{E}[\dot{\ell}(\theta_0, y_i)] = 0$;
- $-\mathbb{E}[\ddot{\ell}(\theta_0, y_i)] = \mathbb{E}[\dot{\ell}(\theta_0, y_i)^2]$.

Those are known as the *Bartlett's identities*.

- First Bartlett Identity:

$$\mathbb{E}\left\{\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right\} = 0, \forall\theta \in \Theta;$$

- Second Bartlett Identity:

$$\mathbb{E}\left\{\left(\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right)^2\right\} = -\mathbb{E}\left\{\frac{d^2}{d\theta^2}\log p_\theta(\boldsymbol{x})\right\}.$$

Then in (2.2) we have

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_0) &= \frac{\dot{\ell}_n(\theta_0, y)/\sqrt{n}}{-\frac{1}{n}\ddot{l}_n(\theta_0, y) - \frac{(\widehat{\theta}_n - \theta_0)}{2n} \cdot \dddot{\ell}_n(\theta^*, y)} \\ &\to_d \frac{Z}{\mathcal{J}(\theta_0)} \\ &\sim N(0, \mathcal{J}(\theta_0)^{-1}),\end{aligned}$$

where $\mathcal{J}$ is known as the Fisher information matrix.

The model might be wrong. But under regularity conditions, the estimator will still converge to the θ' such that your posited model for the response is closest to the true model in KL divergence:

$$KL := \int \log\left(\frac{p_{\text{proposed}}(y)}{p_{\text{true}}(y)}\right) \cdot p_{\text{true}}(y)dy.$$

Assume X is a binary covariate for $i = 1, \cdots, n$. Then

$$\mathbb{P}(Y = 1|X = 1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} = q_1, \quad \mathbb{P}(Y = 1|X = 0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} = q_0,$$

where

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

Thwn we may build the log-likelihood function based on this model:

$$\begin{aligned}
\ell(\beta_0, \beta_1) &= \sum_{i=1}^{n} y_i \log p_i + (1 - y_i) \log(1 - p_i) \\
&= \log(1 - q_1) \cdot \sum_{i=1}^{n} (1 - y_i) x_i + \log(1 - q_0) \cdot \sum_{i=1}^{n} (1 - y_i)(1 - x_i) \\
&= n_{11} \log q_1 + n_{01} \log q_0 + n_{10} \log(1 - q_1) + n_{00} \log(1 - q_0),
\end{aligned}$$

where n_{xy} indicate the number of observations with $X = x, Y = y$. Hence

$$\frac{\partial \ell(q_1, q_0)}{\partial q_j} = \frac{n_{j1}}{q_j} - \frac{n_{j0}}{1 - q_j} \implies \widehat{q}_j = \frac{n_{j1}}{n_{j0} + n_{j1}}.$$

So we have

$$\widehat{\beta}_0 = \log\left(\frac{\widehat{q}_0}{1 - \widehat{q}_0}\right) = \log\left(\frac{n_{01}}{n_{00}}\right).$$

Furthermore, note that

$$\widehat{\beta}_0 + \widehat{\beta}_1 = \log\left(\frac{\widehat{q}_1}{1 - \widehat{q}_1}\right),$$

so

$$\widehat{\beta}_1 = \log\left(\frac{n_{00} n_{11}}{n_{01} n_{10}}\right).$$

3 Exponential Families Model

3.1 Poisson Regression

Apart from logistic regression, another common choice of regression model for discrete outcomes is the Poisson regression. Recall that if $Y \sim Poisson(\lambda)$, then

$$\mathbb{P}(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}, \quad y = 0, 1, \cdots.$$

This model is particularly useful for data where the response are counts without a well-defined upper bound. Assume we have a random sample $Y_1, \cdots, Y_n \overset{\text{i.i.d}}{\sim} Poisson(\lambda)$, then we can derive the likelihood as

$$\begin{aligned}
L_n(\boldsymbol{y}, \lambda) &:= \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \\
\ell_n(\boldsymbol{y}, \lambda) &= \sum_{i=1}^{n} -\lambda + y_i \log \lambda - \log(y_i!) \\
\frac{\partial \ell_n(\boldsymbol{y}, \lambda)}{\partial \lambda} &= -n + \sum_{i=1}^{n} \frac{y_i}{\lambda}.
\end{aligned}$$

Hence the MLE for λ is

$$\widehat{\lambda} = \frac{1}{n}(Y_1 + \cdots + Y_n),$$

i.e., the sample mean. Similar to the logistic regression, we want to decide on how to model the mean λ_i of Poisson response y_i, and the most common choice is that

$$\log \lambda_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + x_{11}\beta_1 + \cdots + x_{1p}\beta_p,$$

which implies $\lambda_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$. We can interpret β_0 as the mean of λ with all other covariates 0 (exclude the intercept), and $\exp^{\beta_j}$ as the proportional increase in the mean of the Poisson random variable (or the rate) for a one unit increase in covariate j while all other covariates are fixed. Under this set up, we may re-write the likelihood function in terms of β_i:

$$\begin{aligned}
\ell_n(\boldsymbol{y}, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) + y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \log(y_i!) \\
\frac{\partial \ell_n(\boldsymbol{y}, \boldsymbol{\beta})}{\partial \beta_i} &= \sum_{j=1}^{n} \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}) \cdot x_{ji} + y_j x_{ji}.
\end{aligned}$$

In this case, solving for the MLE of $\boldsymbol{\beta}$ will require us to solve a system of non-linear equations, which is not possible to obtain a closed form solution. Where one shall use numerical methods like Newton's method, gradient descent, etc..

The advantage of Poisson model is the ability to obtain a mean-variance relationship ($\mathbb{E}[Y] = \mathbb{V}(Y)$) that is not possible under the normal linear model. However, one disadvantage is that the mean equal to variance may not be a realistic assumption in many situations. We could try to extend to allow for variation in the Poisson parameter and then derive the MLE's again, and hence solve for $\boldsymbol{\beta}$.

3.2 Exponential Families

Definition 2 (Linear Exponential Families (LEF))**.** *Suppose $Y_1, \cdots, Y_n \overset{i.i.d}{\sim}$ belongs to the linear exponential family (LEF), then the probability density function (mass function) parametrized by θ is given by*

$$f_{Y_i}(y_i|\theta) = \exp\Big(- b(\theta) + c(y_i) + \theta y_i \Big),$$

where $b(\theta)$ is a function that only depends on the parameter, c is a function that only depends on the data.

We will further assume some regularity conditions hold for the linear exponential family, and throughout the section we will assume the distribution is continuous (discrete case can be easily generalized, we use counting measure rather then Lebesgue measure). For notation simplicity, we will define

$$\ell(y_i, \theta) = \log f_{Y_i}(y_i|\theta), \quad \dot{\ell}(y_i, \theta) = \frac{\partial \ell(y_i, \theta)}{\partial \theta}, \quad \ddot{\ell}(y_i, \theta) = \frac{\partial^2 \ell(y_i, \theta)}{\partial \theta^2},$$

and study some properties of the linear exponential family:

Proposition 1. *Suppose $Y_1 \sim f_{Y_i}(y_i|\theta)$ belongs to the linear exponential family (LEF), then* $\mathbb{E}[\dot{\ell}(y_i, \theta)] = 0$.

Proof. By direct chain rule, we have that

$$\mathbb{E}[\dot{\ell}(y_i, \theta)] = \int_\Theta \frac{\partial \ell(y_i, \theta)}{\partial \theta} \cdot f_{Y_i}(y_i|\theta) dy_i = \int_\Theta \frac{1}{f_{Y_i}(y_i|\theta)} \cdot \frac{\partial f_{Y_i}(y_i|\theta)}{\partial \theta} \cdot f_{y_i}(y_i|\theta) dy_i,$$

we then have

$$\int_\Theta \frac{\partial f_{Y_i}(y_i|\theta)}{\partial \theta} dy_i = \frac{\partial}{\partial \theta} \int_\Theta f_{Y_i}(y_i|\theta) dy_i = \frac{\partial}{\partial \theta} 1 = 0,$$

under regularity conditions.

□

Proposition 2. *Suppose $Y_i \sim f_{Y_i}(y_i|\theta)$ belongs to the linear exponential family (LEF), further assume* $\mathbb{E}[m(y_i, \theta)] = 0$, *then*

$$\mathbb{E}\left[\frac{\partial m(y_i, \theta)}{\partial \theta}\right] = -\mathbb{E}\left[m(y_i, \theta) \cdot \frac{\partial \ell(y_i, \theta)}{\partial \theta}\right].$$

Proof. We have that $\mathbb{E}[m(y_i, \theta)] = 0$, so $d\mathbb{E}[m(y_I, \theta)]/d\theta = 0$. By Chain rule,

$$\begin{aligned} 0 &= \int_\Theta \frac{\partial m(y_i, \theta)}{\partial \theta} \cdot f_{Y_i}(y_i|\theta) dy + \int_\Theta m(y_i, \theta) \cdot \frac{\partial f_{Y_i}(y_i, \theta)}{\partial \theta} \\ &= \mathbb{E}\left[\frac{\partial m(y_i, \theta)}{\partial \theta}\right] + \int_\Theta m(y_i, \theta) \cdot \frac{\partial \ell(y_i, \theta)}{\partial \theta} \cdot f_{Y_i}(y_i|\theta) \\ &= \mathbb{E}\left[\frac{\partial m(y_i, \theta)}{\partial \theta}\right] + \mathbb{E}\left[m(y_i, \theta) \cdot \frac{\partial \ell(y_i, \theta)}{\partial \theta}\right]. \end{aligned}$$

□

Using this proposition, we see that

$$\mathrm{Cov}\left(m(y_i, \theta), \frac{\partial \ell(y_i, \theta)}{\partial \theta}\right) = \mathbb{E}\left[m(y_i, \theta) \cdot \frac{\partial \ell(y_I, \theta)}{\partial \theta}\right] - \mathbb{E}[m(y_i, \theta)] \cdot \mathbb{E}\left[\frac{\partial \ell(y_i, \theta)}{\partial \theta}\right] \equiv 0.$$

The special case of such $m(y_i, \theta)$ is to let $m(y_i, \theta) = \dot{\ell}(y_i, \theta)$, so in this case we have

$$-\mathbb{E}[\ddot{\ell}(y_i, \theta)] = \mathbb{E}[(\dot{\ell}(y_i, \theta))^2],$$

which is also known as the *second Bartlett's identity*. We have shown some useful identities with one single variable, however, the multivariate version can be easily generalized.

We now apply Proposition 1 and 2 to the linear exponential family. Recall for LEF, the probability density takes the form

$$f_{Y_i}(y_i|\theta) = \exp\Big(-b(\theta) + c(y_i) + \theta y_i\Big),$$

or equivalently,

$$\ell(y_i, \theta) = -b(\theta) + c(y_i) + \theta y_i.$$

Now by Proposition 1, $\mathbb{E}[\dot{\ell}(y_i, \theta)] = 0$, so for LEF,

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\Big(-b(\theta) + c(y_i) + \theta y_i\Big)\right] = 0,$$

which implies

$$\mathbb{E}[y_i] = \dot{b}(\theta).$$

Furthermore, by the result in Proposition 2, we have that $-\mathbb{E}[\ddot{\ell}(y_i, \theta)] = \mathbb{E}[(\dot{\ell}(y_i, \theta))^2]$, where

$$\dot{\ell}(y_i, \theta) = -\dot{b}(\theta) + y_i, \quad \ddot{\ell}(y_i, \theta) = -\ddot{b}(\theta).$$

So we have

$$(\dot{\ell}(y_i, \theta))^2 = (y_i - \dot{b}(\theta))^2 = (y_i - \mathbb{E}[y_i])^2 := \mathbb{V}(y_i),$$

which implies

$$\ddot{b}(\theta) = \mathbb{V}(y_i).$$

Example 1. *Consider $Y \sim Poisson(\lambda)$.*

We may write its density function as

$$f_Y(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = \exp\Big(-\lambda + y\log\lambda - \log(y!)\Big).$$

Let $\theta = \log\lambda$, then the density function takes the form

$$f_Y(y|\theta) = \exp\Big(-e^\theta + y\theta - \log(y!)\Big),$$

where the first term is a function of θ only, the second term is linear, and the third term is a function of y only. So Poisson distribution belongs to linear exponential family. If we use the properties of LEF, we have $b(\theta) = e^\theta$, and $\mathbb{E}[Y] = \dot{b}(\theta)$, which means $\mathbb{E}[Y] = e^\theta = \lambda$, and $\mathbb{V}(Y) = \ddot{b}(\theta) = e^\theta = \lambda$, which is true for a Poisson random variable!

Example 2. *Consider $Y \sim Binomial(n, p)$.*

We may write its density function as

$$\begin{aligned} f_Y(y|n,p) &= \binom{n}{y} p^y (1-p)^{n-y} = \exp\left(\log\binom{n}{y} + y\log p + (n-y)\log(1-p)\right) \\ &= \exp\left(\log\binom{n}{y} + y\log\frac{p}{1-p} + n\log(1-p)\right). \end{aligned}$$

Let $\theta = \log\frac{p}{1-p}$, then the density function takes the form

$$f_Y(y|n,\theta) = \exp\left(\log\binom{n}{y} + y\theta - n\log(1+e^\theta)\right),$$

which is, Binomial distribution belongs to linear exponential family, where $b(\theta) = n\log(1+e^\theta)$. Verify yourself that $\mathbb{E}[Y] = \dot{b}(\theta)$, $\mathbb{V}(Y) = \ddot{b}(\theta)$.

3.3 Exponential Dispersal Families

With a small modification to the linear exponential family model, we define the exponential dispersion family models as follows:

Definition 3. *Suppose for a continuous distribution $Y \sim f_Y(y|\theta)$ belongs to the exponential dispersion family, then its density function takes the form*

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

In exponential dispersion family, note that we also have

$$\dot{\ell}(y,\theta) = \frac{y - \dot{b}(\theta)}{a(\phi)} \implies \mathbb{E}\left[\frac{y - \dot{b}(\theta)}{a(\phi)}\right] = 0 \implies \mathbb{E}[Y] = \dot{b}(\theta).$$

As for the variance, we have

$$\mathbb{V}(Y) = a(\phi)\cdot\ddot{b}(\theta) \quad \text{verify it yourself.}$$

Example 3. *Consider $Y \sim N(\mu, \sigma^2)$.*

We may write its density function as

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

where σ^2 is the dispersal parameter.

Example 4. *Suppose $Y \sim Poisson(\lambda)$, where $\lambda \sim Gamma(\alpha(x), \beta(x))$.*

In this case, the marginal distribution of Y is then

$$\begin{aligned} \mathbb{P}(Y = y) &= \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!}\cdot\frac{\lambda^{\alpha(x)-1}}{\Gamma(\alpha(x))\beta(x)^{\alpha(x)}}\cdot\exp\left(-\frac{1}{\beta(x)}\lambda\right)d\lambda \\ &= \frac{1}{y!\Gamma(\alpha(x))\beta(x)^{\alpha(x)}}\int_0^\infty \exp\left(-\left(1+\frac{1}{\beta(x)}\right)\lambda\right)\cdot\lambda^{y+\alpha(x)-1}d\lambda \\ &= \binom{y+\alpha(x)+1}{y}\left(\frac{\beta(x)}{1+\beta(x)}\right)^y\cdot\left(\frac{1}{1+\beta(x)}\right)^{\alpha(x)}, \end{aligned}$$

which is negative binomial.

Jørgensen showed that the standard formulation of the exponential dispersion family for discrete distributions was quite restrictive, there do not exist many discrete distributions which can be expressed as the continuous exponential dispersion family (EDF) distributions. He proposed the following EDF for discrete distributions:

$$f(y,\theta,\phi) = \exp\Big(y\theta - \frac{b(\theta)}{a(\phi)} + c(y,\phi)\Big).$$

In this case, we let $y = x/a(\phi)$ and then we would have the usual exponential dispersion family formulation. Let $\theta = \log\left(\frac{\mu}{\alpha+\mu}\right)$, then the Jørgensen's EDF of the marginal distribution of Y can be expressed as

$$f(y, \mu, \alpha) = \exp\left(y \log \frac{\mu}{\alpha+\mu} + \alpha \log \frac{\alpha}{\alpha+\mu} + \log \binom{y+\alpha-1}{y}\right).$$

3.4 Link Function

Exponential family distributions allow us to work from a very general perspective, although we have only talked about the response so far. In a regression context, we are typically defining the association between Y and some covariates $\boldsymbol{x}$ via the specified probability model. We already know that $\mathbb{E}[Y] = \dot{b}(\theta)$, so we want a model that

$$\dot{b}(\theta) = \mu = \mathbb{E}[y|\boldsymbol{x}].$$

However in this case we might run into some problems. So instead, we will try to connect the covariate vector to the mean through a link function $g(\mu)$, and one approach is to assume that

$$\eta = g(\mu) = \boldsymbol{x}^\top \boldsymbol{\beta}.$$

We choose g such that it is invertible, then g^{-1} is referred as the response function,

$$g^{-1}(\eta) = \mu = \dot{b}(\theta) = \mathbb{E}[y|\boldsymbol{x}].$$

A choice of link function implicitly connects $\boldsymbol{x}$ and β to θ, the natural parameter for the distribution in exponential family from. If $g(\mu) = \theta$, we will refer to g as the canonical link function for that response distribution. Now recall that the log-likelihood of exponential dispersal family can be written as

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi),$$

by choosing the canonical link $g(\mu) = \theta$, we have

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - b(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{a(\phi)} + c(y_i, \phi),$$

where

$$\sum_{i=1}^{n} y_i(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = \sum_{I=1}^{n} y_i \sum_{j=0}^{p} x_{ij}\beta_j = \sum_{j=0}^{p} \beta_j \sum_{i=1}^{n} y_i x_{ij}.$$

Thus the sufficient statistics for β_j's under the canonical link function are

$$\sum_{i=1}^{n} y_i x_{ij},$$

i.e., the sufficient statistics for β_j under the canonical link function are directly related to the covariance of the response with each covariate. To find the MLE for $\boldsymbol{\beta}$, we aim to solve

$$\frac{\partial \ell(\boldsymbol{y}, \boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(y_i, \boldsymbol{\beta})}{\partial \beta_j} = 0.$$

The issue is, the original likelihood is written in terms of θ_i, so to obtain the score functions in terms of β, the following Chain rule is used:

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

We now have

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - \dot{b}(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)};$$

and

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial \dot{b}(\theta_i)}{\partial \theta_i} = \ddot{b}(\theta_i) = \frac{\mathbb{V}(y_i)}{a(\phi)}.$$

Under the linear predictor $\eta_i = \sum_{j=0}^{p} \beta_j x_{ij}$, we have that

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Also, since $\eta_i = g(\mu_i), \mu_i = g^{-1}(\eta_i)$, we have

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \beta_j} &= \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \\
&= \left(\frac{y_i - \mu_i}{a(\phi)}\right) \cdot \left(\frac{a(\phi)}{\mathbb{V}(y_i)}\right) \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij} \\
&= \frac{(y_i - \mu_i)x_{ij}}{\mathbb{V}(y_i)} \times \frac{\partial \mu_i}{\partial \eta_i}.
\end{aligned}$$

Hence by equating the score function to 0, we have

$$\frac{\partial \ell(\boldsymbol{y}, \beta_j)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j}\ell(\boldsymbol{\beta}, \boldsymbol{y}) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mathbb{V}(y_i)} \times \frac{\partial \mu_i}{\partial \eta_i} = 0.$$

Let $\boldsymbol{V}, \boldsymbol{D}$ be diagonal matrices such that

$$\boldsymbol{V} = \operatorname{diag}\Big(\mathbb{V}(y_1), \cdots, \mathbb{V}(y_n)\Big), \boldsymbol{D} = \operatorname{diag}\left(\frac{\partial \mu_1}{\partial \eta_1}, \cdots, \frac{\partial \mu_n}{\partial \eta_n}\right), \tag{3.1}$$

and furthermore let $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ where $\boldsymbol{X}$ is the design matrix, then (3.1) can be written as

$$\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}. \tag{3.2}$$

Equation (3.2) can be viewed as a generalization of the ordinary least squares estimator. In the OLS setting, $\eta_i = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, in this case $\boldsymbol{D} = \boldsymbol{I}$, and under a fixed design we have $\boldsymbol{V} := \sigma^2 \boldsymbol{I}$. Therefore (3.2) under the OLS setting can be written as

$$\boldsymbol{X}^\top \frac{1}{\sigma^2}\boldsymbol{I}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0 \implies \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{y}.$$

3.5 Large Sample Properties of the MLE

Theorem 2. *Under regularity conditions, the MLE $\widehat{\boldsymbol{\beta}}$ converges in distribution to a multivariate normal random variable. i.e.,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathcal{I}_n^{-1}\right),$$

where $\mathcal{I}_n$ is the Fisher information matrix.

For one single observation X_i, we have that

$$\begin{aligned}
(\mathcal{I})_{jk} &= \mathbb{E}\left[-\frac{\partial \ell_i(y_i, \boldsymbol{\beta})}{\partial \beta_j \, \partial \beta_k}\right] \\
&= \mathbb{E}\left[\frac{\partial \ell_i(y_i, \boldsymbol{\beta})}{\partial \beta_j} \cdot \frac{\partial \ell_i(y_i, \boldsymbol{\beta})}{\partial \beta_k}\right] \\
&= \mathbb{E}\left[\frac{(y_i - \mu_i)x_{ij}}{\mathbb{V}(y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{(y_i - \mu_i)x_{ik}}{\mathbb{V}(y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i}\right] \\
&= \frac{x_{ij}x_{ik}}{\mathbb{V}(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.
\end{aligned}$$

Hence the jk-th entry of Fisher's information matrix is given by

$$\mathbb{E}\left[-\frac{\partial^2 \ell(\boldsymbol{y}, \boldsymbol{\beta})}{\partial \beta_j \, \partial \beta_k}\right] = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\mathbb{V}(y_i)} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Let

$$\boldsymbol{W} = \operatorname{diag}\left(\frac{1}{\mathbb{V}(y_i)} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \cdots, \frac{1}{\mathbb{V}(y_n)} \cdot \left(\frac{\partial \mu_n}{\partial \eta_n}\right)^2\right),$$

then it is easy to see that

$$\mathcal{I}_n = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X},$$

where $\boldsymbol{X}$ is the design matrix. Therefore, the large sample property of the MLE states that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1}\right).$$

3.6 Large Sample Properties of the Fitted Values