1 A Brief example for binary response

As mentioned in Chapter 8, given the data $\{(X_i, Y_i)\}_{i=1}^n$, if we want to learn the pattern of a binary response $\pi_i = P(Y_i|X_i)$, we may use Logistic regression:

$$\log \frac{\pi_i}{1-\pi_i} = \beta^T X_i.$$

Although the Logistic Regression is designed for the the binary response, it can be interpreted from a linear model with a logistic noise as following:

$$Y_i = \boldsymbol{I}(\beta^T X_i + \varepsilon_i < 0),$$

where ε_i is a noise drawn from a logistic distribution with zero mean and standard deviation $\frac{\pi}{\sqrt{3}}$. By simple calculus it can be shown that

$$P(Y_i = 1|X_i) = \frac{\exp \beta^T X_i}{1 + \exp \beta^T X_i},$$

which agrees with the form of the Logistic regression. (A logistic distribution with zero mean and standard deviation $\frac{\pi}{\sqrt{3}}$ has the density function $f(x) = \frac{\exp x}{(1+\exp x)^2}$ and the CDF $F(x) = \frac{\exp x}{1+\exp x}$.) Thus, it inspires us that we can combine the indicator function and other noises of any symmetric distribution. For example, if we have $\varepsilon_i \sim \mathcal{N}(0, 1)$, then the Probit regression has the following form:

$$\begin{aligned} Y_i &= \boldsymbol{I}(\beta^T X_i + \varepsilon_i < 0), \\ P(Y_i = 1|X_i) &= \Phi(\beta^T X_i), \end{aligned}$$

where Φ is the CDF of standard normal distribution.

Although we have known how to implement regression for binary case, multi-class case or continuous cases with i.i.d Gaussian noise, but there are still some other categories of data that we haven't dealt with. For example:

1. $Y_i \in \mathbb{N}$, which means the set of possible values for Y_i is a countable infinite collection.

2. $\mathrm{Var}(Y_i) \propto \mathbb{E}Y_i$, which means $\{Y_i\}_{i=1}^n$ may have heteroscedasticity.

How to build an appropriate linear model for the above examples? The answer is the generalized linear models (GLM). Before showing details of GLM, we first introduce the exponential family.

2 Exponential Family

The reason why we consider exponential family is because:

1. It can be shown that, under certain regularity conditions, the exponential family is the only family of distributions with finite-sized sufficient statistics, meaning that we can compress the data into a fixed-sized summary without loss of information. This is particularly useful for online learning, as we will see later.
2. The exponential family is the only family of distributions for which conjugate priors exist, which simplifies the computation of the posterior.
3. The exponential family is at the core of generalized linear models.
4. .....

2.1 Definition

A collection of p.d.f (or p.m.f) $\mathcal{P} = \{p(x|\theta) : x \in \mathcal{X} \text{ and } \theta \in \Omega \subset \mathbb{R}^d\}$ is said to be an **exponential family** if it is of the form

$$\begin{aligned} p(x|\theta) &= h(x) \exp\left[\theta^T \phi(x) - A(\theta)\right], \\ A(\theta) &= \log\left[\int_{\mathcal{X}} h(x) \exp\left(\theta^T \phi(x)\right) d\mu(x)\right]. \end{aligned} \tag{2.1}$$

Here θ is called the natural parameters, $\phi(x) \in \mathbb{R}^d$ is called a vector of sufficient statistics, $A(\theta)$ is called cumulant function, and $h(x)$ is a scaling constant which is often 1. (μ is a measure on $\mathcal{X}$, For example, $\mathcal{X}$ is a Euclidean space and μ is the Lebesgue measure, $\mathcal{X} = \{0, 1\}$ and μ is the counting measure ...)

The equation (2.1) can be generalized to

$$\begin{aligned} p(x|\theta) &= h(x) \exp\left[\eta(\theta)^T \phi(x) - A(\eta(\theta))\right], \\ A(\theta) &= \log\left[\int_{\mathcal{X}} h(x) \exp\left(\eta(\theta)^T \phi(x)\right) d\mu(x)\right]. \end{aligned} \tag{2.2}$$

Then $T = T(x)$ is a **sufficient statistic** for θ if for every t and θ, the conditional distribution of x conditioned on $T = t$, that is $p(x|\theta, T = t)$, does not depend on θ. It means the information of θ is only contained in the statistic T. You may be confused by why we say this. I will give more details later.

Theorem 1 (Factorization Theorem) *A necessary and sufficient condition for a statistic T to be sufficient is that there exist functions $g_\theta \geq 0$ and $h \geq 0$ such that*

$$p(x|\theta) = g_\theta(T(x))h(x), \text{ a.s.}$$

A sufficient statistic T is **minimal sufficient** if for every sufficient statistic $\tilde{T}$ there is a function f such that $T = f(\tilde{T})$ almost surely. A statistic T is called **complete** for $\mathcal{P}$ if

$$\mathbb{E}_\theta[f(T(x))] = c, \text{ for all } \theta,$$

implies $f(T) = c$, almost surely for every $p(x|\theta)$.

An exponential family is called **full rank** if the interior of $\eta(\Omega)$ is not empty and if $\phi(x)$ does not satisfy a linear constraint of the form $\phi(x)^T v = c$ for a non zero constant vector v and a constant c. A exponential family is called a **curved exponential family**, if it satitifies $dim(\theta) < dim(\eta(\theta))$, which is important in sequential experiments. (Examples?). The concept of full-rank exponential family is important because of the following theorems:

Theorem 2 *In a exponential family of full rank, $\phi(x)$ is complete.*

Theorem 3 *If T is complete and sufficient, then it is minimal sufficient. In what situations minimal sufficiency is not completeness?*

A statistic V is called **ancillary** if its distribution does not depend on θ, which means it provides no information about the parameter θ.

Theorem 4 (Basu) *If T is complete and sufficient and V is ancillary, then T and V are independent for under $p(x|\theta)$ for all θ.*

2.2 Properties of cumulant function

$$\nabla_\theta A(\theta) = \mathbb{E}[\phi(x)],$$
$$\nabla^2_\theta A(\theta) = \text{Cov}[\phi(x)].$$

2.3 Examples

2.3.1 Bernoulli

The Bernoulli distribution for $x \in \{0, 1\}$ can be written in the form of exponential family:

$$p(x|\mu) = \mu^x(1-\mu)^{1-x} = \exp\left[x \log \mu + (1-x)\log(1-\mu)\right],$$

where $\phi(x) = (x, 1-x)^T$, $\theta = (\log\mu, \log(1-\mu))$. Notice that $x + (1-x) = 1$, which means $\phi(x)$ is over complete, then we may consider compressing the sufficient statistic to be a minimal sufficient one as following:

$$p(x|\mu) = \mu^x(1-\mu)^{1-x} = \exp\left[x \log(\frac{\mu}{1-\mu}) + \log(1-\mu)\right]$$

where $\phi(x) = x$, $\theta = \log(\frac{\mu}{1-\mu})$ which is called the log-odds ratio.

2.3.2 Univariate Gaussian

The density function of a univariate Gaussian random variable can also be written in the exponential family form:

$$p(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right]$$
$$= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left[-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}\right] \quad = \exp\left[\theta^T\phi(x) - A(\theta)\right]$$

where

$$\theta = (\theta_1,\theta_2)^T = (\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2})^T,$$
$$\phi(x) = (x, x^2)^T,$$
$$A(\theta) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi).$$

2.3.3 Not an example

The uniform distribution family

$$\mathcal{P} = \{p(x|\theta) : p \text{ is a uniform distribution on } [-\theta,\theta],\ \theta \in \mathbb{R}^+\}$$

is not an exponential family.

2.4 More discussions about Completeness and Sufficiency of Exponential Family

2.4.1 Minimal Sufficiency for correlated normals (curved exponential family)

We give an example of the curved family here. In a curved family, there will be some strange phenomenon which we can not find in a full-rank exponential family. And then we will see that minimal sufficiency is not equivalent to Suppose that $(X_i, Y_i), i = 1, \cdots, n$ are sampled i.i.d from the bivariate normal distribution

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}\right)$$

where $\theta \in \Omega = (-1, 1)$. Then we have

$$p(X_i, Y_i, i = 1, \cdots, n|\theta) = \frac{1}{2\pi(1-\theta^2)}\exp\left[\frac{-1}{2(1-\theta^2)}\left(\sum_{i=1}^n (X_i^2 + Y_i^2) - 2\theta\sum_{i=1}^n X_iY_i\right)\right]$$

Let $T = (\sum_{i=1} X_i^2 + Y_i^2, \sum_{i=1}^n X_iY_i)^T$ and $\eta(\theta) = (\frac{-1}{2(1-\theta^2)}, \frac{\theta}{1-\theta^2})^T$, then by the factorization theorem we have T is a sufficient statistic. It is also minimal.

However, it is not a complete one. Let $f(T) = T_1 = \sum_{i=1}^n (X_i^2 + Y_i^2)$, we have $\mathbb{E}[f(T)] = 2n$ is constant, but it doesn't imply $f(T) = 2n$ a.s. The

reason why it is not complete is because the distribution family is not full-rank. The set $\{\eta(\theta) : \theta \in (0,1)\}$ is a hyperbolic curve in $\mathbb{R}^2$, this explained why it is called curved family.

Let $Z_1 = \sum_{i=1}^n X_i^2$, $Z_2 = \sum_{i=1}^n Y_i^2$, we have $Z_1, Z_2 \sim \chi^2(n)$ whatever θ is. Thus, Z_1 and Z_2 are both ancillary statistics. However, (Z_1, Z_2) is not ancillary, because $E[Z_1^2 Z_2^2] = 2n\theta + n^2$, which means (Z_1, Z_2) is not independent of θ.

2.4.2 Bayesian interpretation of sufficiency of exponential family

Assume the prior density for θ is $q(\theta)$ w.r.t the Lebesgue measure. Then given $x \in \mathcal{X}$, the posterior is given as

$$q_{post}(\theta|) = \frac{p(x|\theta)q(x)}{\int_\Omega p(x|\xi)q(\xi)d\xi}$$

Suppose $T(x)$ has the property that, for any prior $q(\cdot)$, the posterior $q_{post}(\cdot)$ depends on x only through $T(x)$. Then we have

$$\begin{aligned} q_{post}(\theta|x) &= \frac{p(x|\theta)q(\theta)}{\int_\Omega p(x|\xi)q(\xi)d\xi} = f(T(x),\theta) \\ p(x|\theta) &= \frac{f(T(x),\theta)}{q(\theta)} \int_\Omega p(x|\xi)q(\xi)d\xi \end{aligned}$$

By the factorization theorem, $T(x)$ is sufficient. Conversely, one can easily show that if $T(x)$ is sufficient for $\mathcal{P}_\theta$, then for any prior $q(\cdot)$, the posterior depends on x only through $T(x)$. Thus, it tells us that why in computations of the posterior of exponential family, we only need to compute the sufficient statistics. That is, all the information we need to perform inference about θ from posterior is the sufficient statistics.

From the viewpoint of Fisher information, we will show why sufficient statistic has no loss of information during data reduction. Let $I^X(\theta)$ be the Fisher information of the data $\{X_1, X_2, \cdots, X_n\}$ and $I^T(\theta)$ be the Fisher information of $T = T(X_1, \cdots, X_n)$. We define the conditional Fisher information of the original data $\{X_1, X_2, \cdots, X_n\}$ conditioned on $T = t$:

$$I^{X|T=t}(\theta) = \mathbb{E}_X \left[\left(\nabla_\theta \log p(X|T=t,\theta)\right) \left(\nabla_\theta \log p(X|T=t,\theta)\right)^T \right]$$

Taking average over t, then we have the conditional Fisher Information

$$I^{X|T}(\theta) = \mathbb{E}_t \left[I^{X|T=t}(\theta) \right]$$

Then we have an equality:

$$I^X(\theta) = I^T(\theta) + I^{X|T}(\theta), \tag{2.3}$$

which means if we the loss of the data reduction $T = T(X_1, \cdots, X_n)$ is

$$\Delta I = I^X(\theta) - I^T(\theta) = I^{X|T}(\theta).$$

If T is sufficient, then $\Delta I = 0$, that is, there is no loss of information.

To see why Fisher information measures the information, we can follow the angle of KL-divergence. Assume θ, θ' are two parameters and they are very closed. Then we have

$$D\left[p(x|\theta), p(x|\theta')\right] \approx \frac{1}{2}(\theta - \theta')^T I^X(\theta)(\theta - \theta').$$

The more convex it is, the more change we have if θ is shifted to a proximal point θ', that is, we can distinguish θ and θ' easier. Thus, it makes sense that Fisher information measures the information the sample containing for the parameter. So far, we have seen the sufficient statistic's importance since it contains all information of the parameters, no matter from a Bayes viewpoint (prior-posterior) or a frequentist viewpoint (Fisher information). That's one of the reasons why exponential family is of interest: because its sufficient statistic is convenient for us to gain and to compute.

2.4.3 Bayes for exponential family

9.2.5 Bayes for the exponential family *

We have seen that exact Bayesian analysis is considerably simplified if the prior is conjugate to the likelihood. Informally this means that the prior $p(\boldsymbol{\theta}|\boldsymbol{\tau})$ has the same form as the likelihood $p(\mathcal{D}|\boldsymbol{\theta})$. For this to make sense, we require that the likelihood have finite sufficient statistics, so that we can write $p(\mathcal{D}|\boldsymbol{\theta}) = p(\mathbf{s}(\mathcal{D})|\boldsymbol{\theta})$. This suggests that the only family of distributions for which conjugate priors exist is the exponential family. We will derive the form of the prior and posterior below.

9.2.5.1 Likelihood

The likelihood of the exponential family is given by

$$p(\mathcal{D}|\boldsymbol{\theta}) \propto g(\boldsymbol{\theta})^N \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T \mathbf{s}_N\right) \tag{9.49}$$

where $\mathbf{s}_N = \sum_{i=1}^N \mathbf{s}(\mathbf{x}_i)$. In terms of the canonical parameters this becomes

$$p(\mathcal{D}|\boldsymbol{\eta}) \propto \exp(N\boldsymbol{\eta}^T\bar{\mathbf{s}} - NA(\boldsymbol{\eta})) \tag{9.50}$$

where $\bar{\mathbf{s}} = \frac{1}{N}\mathbf{s}_N$.

9.2.5.2 Prior

The natural conjugate prior has the form

$$p(\boldsymbol{\theta}|\nu_0, \boldsymbol{\tau}_0) \propto g(\boldsymbol{\theta})^{\nu_0} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{\tau}_0\right) \tag{9.51}$$

Let us write $\boldsymbol{\tau}_0 = \nu_0 \overline{\boldsymbol{\tau}}_0$, to separate out the size of the prior pseudo-data, ν_0, from the mean of the sufficient statistics on this pseudo-data, $\overline{\boldsymbol{\tau}}_0$. In canonical form, the prior becomes

$$p(\boldsymbol{\eta}|\nu_0, \overline{\boldsymbol{\tau}}_0) \propto \exp(\nu_0 \boldsymbol{\eta}^T \overline{\boldsymbol{\tau}}_0 - \nu_0 A(\boldsymbol{\eta})) \tag{9.52}$$

9.2.5.3 Posterior

The posterior is given by

$$p(\boldsymbol{\theta}|\mathcal{D}) = p(\boldsymbol{\theta}|\nu_N, \boldsymbol{\tau}_N) = p(\boldsymbol{\theta}|\nu_0 + N, \boldsymbol{\tau}_0 + \mathbf{s}_N) \tag{9.53}$$

So we see that we just update the hyper-parameters by adding. In canonical form, this becomes

$$p(\boldsymbol{\eta}|\mathcal{D}) \propto \exp\left(\boldsymbol{\eta}^T(\nu_0\overline{\boldsymbol{\tau}}_0 + N\bar{\mathbf{s}}) - (\nu_0 + N)A(\boldsymbol{\eta}))\right) \tag{9.54}$$

$$= p(\boldsymbol{\eta}|\nu_0 + N, \frac{\nu_0\overline{\boldsymbol{\tau}}_0 + N\bar{\mathbf{s}}}{\nu_0 + N}) \tag{9.55}$$

So we see that the posterior hyper-parameters are a convex combination of the prior mean hyper-parameters and the average of the sufficient statistics.

2.4.4 Interpretation of completeness with exponential family examples (Not important in this course, it is just for those who are interested in the origin and the insight of "completeness")

The concept of completeness for a family of measures was introduced in Lehmann and Scheffe (1950) as a precursor to their definition, in the same paper, of a complete statistic. The definition of a complete family lives on only in the (consequently confusingly named) idea of complete statistic (in particular it has nothing to do with the definition of a complete measure defined in measure theory). According to the Lehman Cheffe Theorem, the complete statistics ensures the existence and uniqueness of UMVE.

Theorem 5 *Let $X_1, \cdots, X_n$ be i.i.d samples drawn from $p(x|\theta)$. If $T = T(X_1, \cdots, X_n)$ is a complete and sufficient statistic, there is a function f such that $f(Y)$ is an unbiased estimator of θ, then this function of Y is the unique UMVE of θ.*

The content left following is the origin of complete statistics, I will just have brief overlook but the leave some complex details for readers. We say a family of probability measure $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ on $\mathcal{X}$ is complete if

$$\int_{\mathcal{X}} f(x) dP_\theta(x) = 0, \ \forall \theta \in \Omega, \ \Rightarrow \ P_\theta(\{x : f(x) \neq 0\}) = 0.$$

That is, the family is not complete if there is non-zero function f which is "orthogonal" to every P_θ in some sense. (If you have learned Risez Representation theorem, you can write $\int_{\mathcal{X}} f(x) dP_\theta(x) = \langle f, P_\theta \rangle$ in the sense of dual space.) We will try to gain some intuition for this definition and, thereby, for the definition of a complete statistic.

Assume $\mathcal{X}$ is a finite set, i.e. $\mathcal{X} = \{x_1, \cdots, x_n\}$ for some $n < \infty$. W.L.O.G, we can assume for all $x \in \mathcal{X}$, $p(x|\theta) > 0$ for some θ. (Otherwise we can truncate the set $\mathcal{X}$.) Let $v^\theta = (p(x_1|\theta), \cdots, p(x_n|\theta))^T \in \mathbb{R}^n$.Then $\mathcal{P}$ is complete if and only if $\text{Span}\{v^\theta, \theta \in \Omega\} = \mathbb{R}^n$.

Then we consider a particular exponential family, Poisson distribution family $p(x|\theta) = \frac{\theta^x}{x!} e^{-\theta}$. Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} p(x|\theta)$. Then we have

$$p(x_1, \cdots, x_n|\theta) = \left[\prod_{i=1}^n \frac{1}{x_i!} \right] e^{-n\theta - \log\theta \sum_{i=1}^n x_i}.$$

Let $T = \sum_{i=1}^n X_i$. If $\Omega = \{\theta_1, \cdots, \theta_m\}$ with $1 < m < \infty$, then T is minimal sufficient but not complete. Let's illustrate this by considering the distribution of T.

$$\begin{pmatrix} v^{\theta_1} \\ v^{\theta_2} \\ v^{\theta_3} \\ v^{\theta_4} \\ v^{\theta_5} \end{pmatrix} = \begin{pmatrix} p(T=0|\theta_1) & p(T=1|\theta_1) & \cdots & p(T=K|\theta_1) & \cdots \\ p(T=0|\theta_2) & p(T=1|\theta_2) & \cdots & p(T=K|\theta_2) & \cdots \\ p(T=0|\theta_3) & p(T=1|\theta_3) & \cdots & p(T=K|\theta_3) & \cdots \\ p(T=0|\theta_4) & p(T=1|\theta_4) & \cdots & p(T=K|\theta_4) & \cdots \\ p(T=0|\theta_5) & p(T=1|\theta_5) & \cdots & p(T=K|\theta_5) & \cdots \end{pmatrix}_{5\times\infty} = \boldsymbol{A}$$

which can be seen a matrix whose rows are five vectors in a $\ell^2(\mathbb{N})$. Thus, we can find a $\boldsymbol{w} = (w_1, w_2, \cdots)$ such that $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{0}$. Let $f(k) = w_k$, then $E_{\theta_k} f(T) = 0$ for all $k = 1, 2, 3, 4, 5$. But it is not a zero function.

If we take $\Omega = \{0, \pi, 2\pi, \cdots, k\pi, \cdots\}$, T is also minimal sufficient but not complete. The construction of f is related to Taylor expansion. We define $f(T)$ as follows:

$$f(T) = \begin{cases} (-1)^{\frac{T+1}{2}} & T \text{ is an odd number} \\ 0 & T \text{ is an even number} \end{cases}$$

However, if Ω is an infinite set and $\inf \Omega = 0$, then T is minimal sufficient and complete. That is because if if it has a sequence of θ converging to 0 such that $E_\theta f(T) = 0$, then by some property of analytic functions, for all θ in an interval near 0 this property still holds.

3 Generalized Linear Models

3.1 Definition

Linear and logistic regression are examples of generalized linear models, or GLMs. These are models in which the output density is in the exponential family and in which the mean parameters are a linear combination of the inputs, passed through a possibly nonlinear function, such as the logistic function. We describe GLMs in more detail below. We focus on scalar outputs for notational simplicity.

Consider the following distribution

$$p(y_i|\theta_i, \sigma^2) = \exp\left[\frac{y_i\theta_i - A(\theta_i)}{\sigma^2} + c(y_i, \sigma^2)\right],$$

where σ^2 is called the dispersion parameter, θ_i is the natural parameter, A is the cumulant function and c is normalized term. For example, in Logistic regression, $\theta_i = \log \frac{\mu_i}{1-\mu_i} = \boldsymbol{w}^T X_i$ where $\mu_i = \mathbb{E}Y_i$.

In this case, we can write $\theta_i = g(\mu_i)$ for $g(\mu_i) = \log \frac{\mu_i}{1-\mu_i}$. The function $g(\cdot)$ is called link function, which is often monotonic. We are free to choose almost any function we like for $g(\cdot)$, so long as it is invertible, and so long as $g^{-1}(\cdot)$ has the appropriate range. The function g^{-1}, which satisfies $\mu_i = g^{-1}(\boldsymbol{w}^T X_i)$, is sometimes called mean function.

If y_i can be written in the form of exponential family and $\theta_i = \boldsymbol{w}^T X_i$ is the natural parameter, i.e.

$$p(y_i|\boldsymbol{w}, X_i, \sigma^2) = \exp\left[\frac{y_i\boldsymbol{w}^T X_i - A(w^T X_i)}{\sigma^2} + c(y_i, \sigma^2)\right],$$

then g is called canonical link function (for example logistic regression and the linear regression).

We then give some simple examples:

- For linear regression, we have

$$\log p(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) = \frac{y_i\mu_i - \frac{\mu_i^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2)\right) \quad (9.83)$$

  where $y_i \in \mathbb{R}$, and $\theta_i = \mu_i = \mathbf{w}^T\mathbf{x}_i$ Here $A(\theta) = \theta^2/2$, so $\mathbb{E}[y_i] = \mu_i$ and $\text{var}[y_i] = \sigma^2$.

- For binomial regression, we have

$$\log p(y_i|\mathbf{x}_i, \mathbf{w}) = y_i \log(\frac{\pi_i}{1-\pi_i}) + N_i \log(1-\pi_i) + \log \binom{N_i}{y_i} \tag{9.84}$$

where $y_i \in \{0, 1, \ldots, N_i\}$, $\pi_i = \text{sigm}(\mathbf{w}^T \mathbf{x}_i)$, $\theta_i = \log(\pi_i/(1-\pi_i)) = \mathbf{w}^T \mathbf{x}_i$, and $\sigma^2 = 1$. Here $A(\theta) = N_i \log(1 + e^\theta)$, so $\mathbb{E}[y_i] = N_i \pi_i = \mu_i$, $\text{var}[y_i] = N_i \pi_i (1-\pi_i)$.

- For **poisson regression**, we have

$$\log p(y_i|\mathbf{x}_i, \mathbf{w}) = y_i \log \mu_i - \mu_i - \log(y_i!) \tag{9.85}$$

where $y_i \in \{0, 1, 2, \ldots\}$, $\mu_i = \exp(\mathbf{w}^T \mathbf{x}_i)$, $\theta_i = \log(\mu_i) = \mathbf{w}^T \mathbf{x}_i$, and $\sigma^2 = 1$. Here $A(\theta) = e^\theta$, so $\mathbb{E}[y_i] = \text{var}[y_i] = \mu_i$. Poisson regression is widely used in bio-statistical applications, where y_i might represent the number of diseases of a given person or place, or the number of reads at a genomic location in a high-throughput sequencing context (see e.g., (Kuan et al. 2009)).

3.2 Probit Regression

In binary response, we use Logistic Regression $P(Y_i = 1|X_i, w) = \text{sigm}(\boldsymbol{w}^T X_i)$. Here we consider a new type of link function Φ which is the CDF for standard Gaussian. This is known as Probit regression. We then can easily find the MLE for probit regression using gradient methods. Let $z_i = \boldsymbol{w}^T X_i$ and $\tilde{y}_i \in \{-1, +1\}$. Then the gradient of the log-likelihood the i-th sample is :

$$\frac{d}{d\boldsymbol{w}} \log p(\tilde{y}_i|\boldsymbol{w}^T X_i) = \frac{dz_i}{dw} \frac{d}{dz_i} \log p(\tilde{y}_i|\boldsymbol{w}^T X_i) = X_i \frac{\tilde{y}_i \phi(z_i)}{\Phi(\tilde{y}_i z_i)}$$

where ϕ is the p.d.f for standard Gaussian and Φ is the c.d.f for standard Gaussian. Similarly, we can derive the Hessian matrix for the log-likelihood function of i-th sample

$$\boldsymbol{H}_i = -X_i \left(\frac{\phi(z_i)^2}{\Phi(\tilde{y}_i z_i)^2} + \frac{\tilde{y}_i z_i \phi(z_i)}{\Phi(\tilde{y}_i z_i)} \right) X_i^T.$$

3.3 Newton-Raphson and Fisher Scoring

The algorithm to solve the Probit regression estimators is called Newton-Raphson algorithm. Actually, it can be used to solve the M.L.E for generalized linear models. Let $\ell(\boldsymbol{w})$ be the log-likelihood. We denote $\boldsymbol{u}(\boldsymbol{w})$ and $\boldsymbol{H}$ as its gradient and Hessian matrix, respectively. Then, the M.L.E is aimed to solve the following equation:

$$\mathbf{0} = \boldsymbol{u}(\hat{\boldsymbol{w}}).$$

To solve the above, we adopt the Newton-Raphson methods and update the parameter iteratively:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \boldsymbol{H}^{-1}(\boldsymbol{w}^{(t)}) \boldsymbol{u}(\boldsymbol{w}^{(t)}).$$

Sometimes we can accelerate the above iteration by adopting the Fisher's Scoring methods:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \boldsymbol{I}^{-1}(w^{(t)})\boldsymbol{u}(w^{(t)}).$$

where $\boldsymbol{I}(\boldsymbol{w})$ is the Fisher information matrix at $\boldsymbol{w}$.

In a GLM, we have $\ell = \sum_{i=1}^{n} \ell_i$, where

$$\ell_i = \left[y_i\theta_i - A(\theta_i)\right]/\sigma^2 + c(y_i, \sigma^2),$$

and the link function $g(\cdot)$ links $\mu_i = Ey_i = A'(\theta_i)$ and $\eta_i = \boldsymbol{w}^T X_i$ via $\eta_i = g(\mu_i)$. Denote $\boldsymbol{w} = (w_1, \cdots, w_p)^T$ and $X_i = (x_{i1}, \cdots, x_{ip})^T$, we need to calculate

$$u_{ir} = \frac{\partial \ell_i}{w_r} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial w_r}.$$

Note that

$$\begin{aligned}
&\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - \mu_i}{\sigma^2}, \qquad \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{A''(\theta_i)},\\
&\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}, \qquad \frac{\partial \eta_i}{\partial w_r} = x_{ir},
\end{aligned}$$

we combine the equalities above and yield that

$$u_{ir} = \frac{(y_i - \mu_i)x_{ir}}{\sigma^2 A''(\theta_i)g'(\mu_i)}$$

To make it clear, we write $\boldsymbol{u}(\boldsymbol{w}) = (u_1, \cdots, u_p)^T$, where $u_i = \sum_{r=1}^{n} u_{ir}$. It follows that the (r, s)-entry of $\boldsymbol{I}(\boldsymbol{w})$ is

$$\mathrm{Cov}(u_r, u_s) = \sum_{i=1}^{n} \frac{\mathrm{Var}(y_i)x_{ir}x_{is}}{\left(\sigma^2 A''(\theta_i)\right)^2 g'(\mu_i)^2} = \sum_{i=1}^{n} \frac{x_{ir}x_{is}}{\sigma^2 A''(\theta_i)g'(\mu_i)^2}$$

Let $\boldsymbol{X}$ be the design matrix where $\boldsymbol{X}_{ir} = x_{ir}$ and the weighted matrix $\boldsymbol{W}$ be a diagonal matrix with main diagonal elements $w_{ii} = \frac{1}{\sigma^2 A''(\theta_i)g'(\mu_i)^2}$. It yields that

$$\boldsymbol{I}(\boldsymbol{w}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}.$$

Since $u_{ir} = w_{ii}(y_i - \mu_i)g'(\mu_i)x_{ir}$, we have

$$u_r = (\boldsymbol{Y} - \boldsymbol{\mu})^T\boldsymbol{W}\boldsymbol{G}'(\boldsymbol{\mu})\boldsymbol{X}^{(r)},$$

where $\boldsymbol{Y} = (y_1, \cdots, y_n)^T$, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)^T$, $\boldsymbol{G}'(\boldsymbol{\mu})$ is the diagonal matrix with diagonal elements $\boldsymbol{G}'(\boldsymbol{\mu})_{(ii)} = g'(\mu_i)$ and $\boldsymbol{X}^{(r)} = (x_{1r}, x_{2r}, \cdots, x_{nr})^T$ is

the r-th column vector of $\boldsymbol{X}$. Thus, by the commutativity of the diagonal matrices $\boldsymbol{W}$ and $\boldsymbol{G}'(\boldsymbol{\mu})$, we have

$$\boldsymbol{u}(\boldsymbol{w}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{G}'(\boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu}),$$

One can see that

$$\frac{\partial(\boldsymbol{y} - \boldsymbol{\mu})}{\partial \boldsymbol{w}^T} = -\boldsymbol{G}'(\boldsymbol{\mu})^{-1} \boldsymbol{X},$$

so

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{w}) &= \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{G}'(\boldsymbol{\mu}) \frac{\partial(\boldsymbol{Y} - \boldsymbol{\mu})}{\partial \boldsymbol{w}^T} + \frac{\partial \left[\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{G}'(\boldsymbol{\mu}) \right]}{\partial \boldsymbol{w}^T} (\boldsymbol{Y} - \boldsymbol{\mu}) \\ &= -\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} + \boldsymbol{X}^T \boldsymbol{D} \frac{\partial \left[\boldsymbol{W} \boldsymbol{G}'(\boldsymbol{\mu}) \mathbf{1}_p \right]}{\partial \boldsymbol{w}^T}, \end{aligned}$$

where $\mathbf{1}_p = (1, 1, \cdots, 1)$ is a p-dimensional vector with all elements taking 1 and $\boldsymbol{D}$ denotes the diagonal matrix with diagonal elements $\boldsymbol{D}_{ii} = y_i - \mu_i$, The above equation implies that

$$\mathbb{E}_{\boldsymbol{w}} \left[\boldsymbol{H}(\boldsymbol{w}) \right] + \boldsymbol{I}(\boldsymbol{w}) = \boldsymbol{X}^T (\mathbb{E}\boldsymbol{D}) \frac{\partial \left[\boldsymbol{W} \boldsymbol{G}'(\boldsymbol{\mu}) \mathbf{1}_p \right]}{\partial \boldsymbol{w}^T} = \mathbf{0}_{p \times p}.$$

Thus, it is partly due to the simpler form of $\boldsymbol{I}(w)$ than $\boldsymbol{H}(\boldsymbol{w})$, people prefer to use the Fisher's Scoring method than the Newton-Raphson method. But these two methods coincide if the link-function $g(\cdot)$ is a canonical link. To see this, we have $\eta_i = \theta_i$ and

$$u_{ir} = \frac{\partial \ell_i}{w_r} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial w_r} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \eta_i}{\partial w_r} = \frac{(y_i - \mu_i) x_{ir}}{\sigma^2}.$$

which means that $\boldsymbol{W} \boldsymbol{G}'(\boldsymbol{\mu}) = \frac{1}{\sigma^2} \boldsymbol{I}_{p \times p}$ where $\boldsymbol{I}_{p \times p}$ is a $p \times p$ identity matrix and is independent of $\boldsymbol{w}$. Thus, the second term of $\boldsymbol{H}(\boldsymbol{w})$ is always $\mathbf{0}_{p \times p}$.

The Fisher's Scoring method is also related to weighted least square in some sense. The Fisher's Scoring method iterates as following:

$$\begin{aligned} \boldsymbol{w}^{(t+1)} &= \boldsymbol{w}^{(t)} + \boldsymbol{I}^{-1}(\boldsymbol{w}^{(t)}) \boldsymbol{u}(\boldsymbol{w}^{(t)}) \\ &= \left(\boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{z}^{(t)} \end{aligned}$$

where $\boldsymbol{z}^{(t)} = \boldsymbol{X} \boldsymbol{w}^{(t)} + \boldsymbol{G}'(\boldsymbol{\mu}^{(t)})(\boldsymbol{Y} - \boldsymbol{\mu}^{(t)}) = g(\boldsymbol{\mu}^{(t)}) + \boldsymbol{G}'(\boldsymbol{\mu}^{(t)})(\boldsymbol{Y} - \boldsymbol{\mu}^{(t)})$. If pretending that $\boldsymbol{W}^{(t)}$ and $\boldsymbol{z}^{(t)}$ is independent of $\boldsymbol{w}^{(t)}$, $w^{(t+1)}$ is a solution to

$$\boldsymbol{w}^{(t+1)} = \arg \min_{\boldsymbol{w} \in \mathbb{R}^p} \left(\boldsymbol{z}^{(t)} - \boldsymbol{w} \right)^T \boldsymbol{W}^{(t)} \left(\boldsymbol{z}^{(t)} - \boldsymbol{w} \right).$$

3.4 Generalized linear mixed model (GLMM)

LMM is a famous model, as an extension of linear models with random effects. We now introduce GLMM.

$$p(y_i|u, w, \sigma^2) = \exp\left[(y_i\theta_i)/\sigma^2 + c(y_i, \sigma^2)\right], \text{ where } \boldsymbol{u} \sim f(\boldsymbol{u}|\boldsymbol{D}).$$

where $\theta_i = \boldsymbol{w}^T X_i + \boldsymbol{u}^T Z_i$, with X_i being the i-th row of $\boldsymbol{X}$, the model matrix for fixed effects, and likewise with Z_i the i-th row of $\boldsymbol{Z}$ the matrix of random effects. Then the likelihood is given as

$$L(w, \sigma^2, \boldsymbol{D}|Y) = \int \prod_{i=1}^{n} p(y_i|u, w, \sigma^2) f(u|\boldsymbol{D}) du,$$

which cannot usually be evaluated in closed form and has an integral with dimension equal to the number of levels of the random factors $\boldsymbol{u}$. To solve this problem, we can use Monte Carlo EM algorithm, Monte Carlo Newton-Raphson algorithm or Simulated Maximum Likelihood algorithm.