Generalized Linear Models: Logistic Regression and Beyond

Authors: M. Mattheakis, P. Protopapas

1 Introduction

Ordinary Linear regression is a simple and well studied model of statistical learning. Despite its simplicity, this model has been successfully applied in a wide range of real-world applications. Nevertheless, there are plenty of situations where the simple linear regression model fails. The linear regression model assumes that the observations are obtained by a Normal distribution with mean that linearly depends on predictors, however, this assumption is not satisfied in many problems. For instance, many real-world observations are binary, such as data that consists of "yes" or "no" responses. In this case we could use Bernoulli distribution or, more general, bionomial distribution leading to the *Logistic regression* model. Furthermore, there are many times that the observations only occur on the positive real axis rather than the entirety of the reals. For such situations we would use exponential or gamma distributions for the observations instead of Normal distribution. That necessitates and inspires us to develop a more flexible and general approach in the context of *generalized linear models* (GLMs). The formulation of GLMs is based on the generalization of two fundamental assumptions of the linear regression. On the contrary to linear regression model, GLMs do not require a linear relationship between the expectation value and the predictors and do not assume Normal distribution for the error term.

In these notes, we introduce the idea and develop the theory of GLMs. In this general framework, the observations can be integer-valued, non-negative, categorical, or otherwise unsatisfactory for a simple linear model. The critical point here is that, although the observations can be unsatisfactory for a linear model, we can perform a transformation to the expectation value that is linear to the predictors and thus, we retain the linear relationship. In section 2.1, we start with a brief overview of the *linear regression* approach. The formulation of GLMs is shown in two generalizations of the simple linear regression model and presented in sections 2.2 and 2.3. In particular, in section 2.2 we perform the first generalization of the linear regression model where we investigate the general case that the observations are distributed about a linear predictor with a distribution that belongs to the exponential family. Normal, Bernoulli, binomial, Poisson, exponential, gamma, and negative binomial distributions are special cases of the exponential family. In section 2.3 we make the second generalization in the simple linear regression model that leads to GLM. In particular, we introduce the *Link function* that transforms the means to be linear with the predictors. *Linear and Logistic* regression models, which are special

cases of the GLMs, are presented in the end of this section as special examples. Finally, in section 3, we discuss the *maximum likelihood estimation* in the overall framework of GLMs for canonical links (section 3.1) and for general links (section 3.2).

2 Generalized Linear Models

In this section, we formulate the *generalized linear models* (GLMs) approach by performing two generalizations in the linear regression model. As examples, we derive the linear and logistic regression models in the context of the general GLM framework.

2.1 Linear regression

Linear regression is a simple approach for supervised learning for predicting a quantitative response variable. Although linear regression is a straightforward model, it is a still useful and widely used statistical learning method. In addition, linear regression serves as a good jumping-off point for newer and more flexible approaches such as GLMs. In this section, we give a brief overview of the foundations of linear regression.

We assume a training dataset with n training data-points $\{y_i, \mathbf{x}_i\}$ (with $i = 1, ..., n$), where each pair consists of an one dimensional response variable $y_i \in \mathbb{R}$, and a $(p+1)$ dimensional input (predictor) vector $\mathbf{x}_i \in \mathbb{R}^{p+1}$, where p indicates the number of the predictors. In a regression model we aim to find a relationship between the quantitative response y_i, or in matrix representation $Y \equiv (y_1, ..., y_n)^T$, on the basis of predictor variables matrix $\mathbf{X} \equiv (\mathbf{x}_1^T, ..., \mathbf{x}_n^T)^T$ of the form

$$Y = f(\mathbf{X}) + \epsilon, \tag{1}$$

where f is some fixed but unknown function of $\mathbf{X}$, and $\epsilon \equiv (\epsilon_1, ..., \epsilon_n)^T$ is a random *error term* (or stochastic noise) which is considered independent on $\mathbf{X}$. The matrix $\mathbf{X}$ is called the *design matrix* and essentially it is a matrix of row-vectors $\mathbf{x}_i$ defined as

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_{11} & \cdots & \mathbf{x}_{1p} \\ 1 & \mathbf{x}_{21} & \cdots & \mathbf{x}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{x}_{n1} & \cdots & \mathbf{x}_{np} \end{pmatrix}. \tag{2}$$

We observe that there is an 0^{th} row in $\mathbf{X}$, i.e. $\mathbf{x}_{i0} = 1$ $(i = 1, ..., n)$, that explains the $(p + 1)$ dimensions of $\mathbf{x}_i$ vectors; we will discuss the role of this row in a while.

There are two fundamental assumptions in the context of linear regression. The first assumption states that there is approximately a linear relationship between $\mathbf{X}$ and Y, in other words, a linear relationship between the expected value $\mathbb{E}[y_i]$ and the predictors $\mathbf{x}_i$. The second assumption states that each observation y_i is independently distributed about the linear predictors with a *Normal distribution with zero mean*, that is, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, where $\mathcal{N}$ denotes the Normal distribution and σ^2 is the variance. Mathematically, by using the two above fundamental assumptions of linear regression, the formula (1) is written as the

linear relationship:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \quad \text{or} \tag{3}$$

$$Y = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{4}$$

where $\boldsymbol{\beta} \equiv (\beta_0,\ \beta_1, ..., \beta_p)^T$ ($\boldsymbol{\beta} \in \mathbb{R}^{p+1}$) is a vector of coefficients that will be estimated by the likelihood maximization (see advanced section 2), $\mathbf{x}_i^T\boldsymbol{\beta}$ is the dot product

$$\mathbf{x}_i^T \boldsymbol{\beta} = \sum_{p=0}^{p} x_{ip}\beta_p, \tag{5}$$

and $\mathbf{X}\boldsymbol{\beta}$ is a matrix product. We notice that there is a 0^{th} element in the vector $\boldsymbol{\beta}$, namely β_0, which is called the *intercept* and corresponds to the constant unity row $\mathbf{x}_{i0}$ of the design matrix $\mathbf{X}$. This term captures the bias in the linear regression model. The intercept term is required in many statistical inference procedures for linear models, however, in theoretical considerations β_0 is often suggested to be zero.

In linear regression model, the expectation value μ_i (first moment) is linearly dependent on the predictors, as

$$\mu_i = \mathbb{E}\left[y_i\right] = \mathbf{x}_i^T\boldsymbol{\beta}, \tag{6}$$

and the variance is

$$\text{var}\left[y_i\right] = \mathbb{E}\left[(y_i - \mu_i)^2\right] = \sigma^2, \tag{7}$$

which can be equivalently written as the conditional Normal distribution of y_i on $\mathbf{x}_i$ as

$$p(y_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2) = \mathcal{N}(\mu_i, \sigma^2). \tag{8}$$

The formulation of the GLMs essentially demands the relaxation and generalization of the two aforementioned assumptions of the linear regression model. Firstly, for the *random component* we generalize the error distribution (8) by using the general *canonical exponential family* instead of the the Normal distribution, which is included as a special case, thus,

$$p(y_i|\mathbf{x}_i) = \text{Canonical Exponential Family}. \tag{9}$$

Secondly, we generalize the *systematic components* of the model, that is, the linear relation (6) between μ_i and $\mathbf{x}_i$, by introducing the *Link function* $g(\mu_i)$ which transforms μ_i to be linear with the predictors $\mathbf{x}_i$, hence,

$$g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}, \tag{10}$$

where in the linear regression model g is the identity function. These two generalizations yield GLMs formulation providing a flexible and efficient statistical learning method.

2.2 The Canonical Exponential Family

In this section we perform the first generalization in the linear regression model which is required for the formulation of GLMs. In particular, we generalize the distribution of

errors, that is, from the Normal distribution to the more general *canonical exponential family* distributions. The exponential family is a pretty wide range of distributions that includes special cases like many of the known distributions such as Normal, Bernoulli, binomial, Poisson, exponential, gamma, inverse Gaussian, and negative binomial distributions. The *probability density* (or *probability mass function*) of the canonical exponential family is

$$f_\theta(y) = f(y|\theta) = \exp\left(\frac{y\,\theta - b(\theta)}{\phi} + c(y,\phi)\right), \tag{11}$$

where y is the dependent variable, θ and ϕ are parameters, and $b(\theta)$ and $c(y,\phi)$ are known functions determined by the distribution. More specifically, θ is called *canonical* or *natural* parameter of the distribution and is the parameter of interest, ϕ is a scale parameter called *dispersion parameter* and related to the variance, $b(\theta)$ depends only on θ (not on y) and completely characterizes the distribution, subsequently it is the *cumulant function*, and $c(y,\phi)$ is the normalization factor.

In regression modeling situations the distribution of each y_i varies by the observation through the subscript i. It is customary to let the distribution family remain constant, in our case the exponential distribution family, but allow the canonical and dispersion parameters to vary by observation by the notation θ_i and ϕ_i, respectively, where the dispersion parameter is determined by the *prior weights* w_i. In particular, when the dispersion parameter varies by observation, it is according to $\phi_i = \phi/w_i$, that is, a constant divided by known weight factors w_i. When each pair of the observations has different dispersion parameter ϕ_i we have *heteroskedasticity*, otherwise when $\phi_i = \phi$ we have *homoskedasticity*. We assume that the observations y_i are independent of each other and given by the distribution density

$$f_{\theta_i}(y_i) = \exp\left(\frac{y_i\,\theta_i - b(\theta_i)}{\phi_i} + c(y_i,\phi_i)\right) \tag{12}$$

From the density of Eq. (12) and the independence among observations, the Likelihood is

$$L(y_i|\theta_i) = \prod_{i=1}^{n} f_{\theta_i}(y_i). \tag{13}$$

The *log-likelihood* is the fundamental quantity for the statistical inference which is defined as:

$$\ell(y_i|\theta_i) = \log L(y_i|\theta_i) = \sum_{i=1}^{n} \log f_{\theta_i}(y_i), \tag{14}$$

where log here is the natural logarithm. In general, it is easier and numerically more stable to work with the log-likelihood instead of the likelihood, since the product turns to summation. Two important identities regarding the log-likelihood concern the *score function*

$$S(\theta_i) = \frac{\partial \ell(y_i|\theta_i)}{\partial \theta_i}, \tag{15}$$

where the maximum likelihood estimator determines the root $\tilde{\theta}_i$ of the score function, $S(\tilde{\theta}_i = 0)$. The two identities regarding the score function, and in turn the log-likelihood,

read

$$\mathbb{E}\left[S(\theta_i)\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right] = 0 \tag{16}$$

and

$$I(\theta_i) \equiv -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial {\theta_i}^2}\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right]^2 = \mathrm{var}(S(\theta_i)), \tag{17}$$

where $I(\theta_i)$ is the *Fisher information matrix* and $\mathbb{E}\left[.\right]^\nu$ denotes the ν moment. For the proof of the identities (16) and (17) we use the following relations:
The score function can be written as

$$S(\theta_i) = \frac{\partial \ell}{\partial \theta_i} = \frac{1}{f_{\theta_i}} \frac{\partial f_{\theta_i}}{\partial \theta_i}, \tag{18}$$

the second derivative of the log-likelihood is

$$\frac{\partial^2 \ell}{\partial {\theta_i}^2} = \frac{1}{{f_{\theta_i}}^2} \left(f_{\theta_i} \frac{\partial^2 f_{\theta_i}}{\partial {\theta_i}^2} - \left(\frac{\partial f_{\theta_i}}{\partial \theta_i} \right)^2 \right), \tag{19}$$

the ν moment of an arbitrary function h in the distribution $f_{\theta_i}(y_i)$ is given by

$$\mathbb{E}\left[h\right]^\nu = \int_{y_i} h^\nu f_{\theta_i}(y_i) dy_i. \tag{20}$$

Due to the fact that y_i are independent of each other the variance is defined by

$$\mathrm{var}\left[h\right] = \mathbb{E}\left[\left(h - \mathbb{E}\left[h\right]\right)^2\right] = \mathbb{E}\left[h^2\right] - \mathbb{E}\left[h\right]^2, \tag{21}$$

and finally for a well defined probability density we require

$$\int_{y_i} f_{\theta_i}(y_i)\, dy_i = 1. \tag{22}$$

Proof of identity (16):

$$\mathbb{E}\left[S(\theta_i)\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right] = \int_{y_i} \frac{1}{f_{\theta_i}} \frac{\partial f_{\theta_i}}{\partial \theta_i} f_{\theta_i} dy_i = \frac{\partial}{\partial \theta_i} \int_{y_i} f_{\theta_i}\, dy_i = 0,$$

where we use the regularity condition to take the derivative out of the integral. ■
Proof of identity (17):

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial {\theta_i}^2}\right] = \mathbb{E}\left[\frac{1}{f_{\theta_i}} \frac{\partial^2 f_{\theta_i}}{\partial {\theta_i}^2}\right] - \mathbb{E}\left[\left(\frac{1}{f_{\theta_i}} \frac{\partial f_{\theta_i}}{\partial \theta_i}\right)^2\right],$$

where the first term in the right-hand is zero:

$$\mathbb{E}\left[\frac{1}{f_{\theta_i}} \frac{\partial^2 f_{\theta_i}}{\partial {\theta_i}^2}\right] = \int_{y_i} \frac{1}{f_{\theta_i}} \frac{\partial^2 f_{\theta_i}}{\partial {\theta_i}^2} f_{\theta_i} dy_i = \frac{\partial^2}{\partial {\theta_i}^2} \int_{y_i} f_{\theta_i} dy_i = 0,$$

while the second term in the right-hand reads

$$\mathbb{E}\left[\left(\frac{1}{f_{\theta_i}}\frac{\partial f_{\theta_i}}{\partial \theta_i}\right)^2\right] = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta_i}\right)^2\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right]^2 = \text{var}(S(\theta_i)),$$

since $\mathbb{E}\left[\partial \ell / \partial \theta_i\right] = 0$ the second moment gives the variance. Subsequently, we prove that

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial {\theta_i}^2}\right] = -\mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right]^2 . \blacksquare$$

Having in our disposal the identities (16) and (17) we can derive the general formulas for the mean and variance in the exponential family distributions:

$$\mu_i = \mathbb{E}\left[y_i\right] = b'(\theta_i), \tag{23}$$

and

$$\text{var}\left[y_i\right] = \mathbb{E}\left[\left(y_i - \mu_i\right)^2\right] = \phi_i b''(\theta_i), \tag{24}$$

where primes denote derivatives with respect to θ_i. Hence, both the mean and the variance are functions of the canonical parameter θ_i. From Eqs. (23) and (24) we read that $b(\theta_i)$ is the *cumulant function* of the distribution, since it completely determines the first two moments. In addition, the expression (24) states that for a given positive ϕ_i, which is usually true, the cumulant function is strictly concave ($b(\theta_i) > 0$) since its second derivative is always positive (variance is positive by definition). Furthermore, using the prior weights w_i ($\phi_i = \phi / w_i$), Eq. (24) is written as $\text{var}\left[y_i\right] = \phi b''(\theta_i)/w_i$ and states that a larger weight w_i implies a smaller variance. For the proof of the relations (23) and (24) we use the following expressions for the derivatives of log-likelihood of the exponential density:

$$\ell = \log f_{\theta_i} = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{\phi_i} + \sum_{i=1}^{n} c(y_i, \phi_i), \tag{25}$$

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\phi_i}, \tag{26}$$

and

$$\frac{\partial^2 \ell}{\partial {\theta_i}^2} = -\sum_{i=1}^{n} \frac{b''(\theta_i)}{\phi_i}. \tag{27}$$

Proof of Eq. (23):
Starting from the identity (16),

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right] = \mathbb{E}\left[\sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\phi_i}\right] &= \sum_{i=1}^{n} \mathbb{E}\left[\frac{y_i - b'(\theta_i)}{\phi_i}\right] \\
&= \sum_{i=1}^{n} \frac{1}{\phi_i} \mathbb{E}\left[y_i\right] - \sum_{i=1}^{n} \frac{1}{\phi_i} \mathbb{E}\left[b'(\theta_i)\right] = 0 \\
&\Rightarrow \mu_i \equiv \mathbb{E}\left[y_i\right] = b'(\theta_i),
\end{aligned}$$

we note that the expectation value is defined by an integral over y_i and thus, the summation can be taken out of the $\mathbb{E}[.]$. Moreover, b depends only on θ_i hence, $\mathbb{E}[b'(\theta_i)] = b'(\theta_i)$. ■

Proof of Eq. (24):
Starting from identity (17),

$$\mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta_i}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial {\theta_i}^2}\right] \Rightarrow \mathbb{E}\left[\left(\sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\phi_i}\right)^2\right] = -\mathbb{E}\left[-\sum_{i=1}^{n} \frac{b''(\theta_i)}{\phi_i}\right]$$
$$\Rightarrow \sum_{i=1}^{n} \frac{1}{{\phi_i}^2} \mathbb{E}\left[(y_i - \mu_i)^2\right] = \sum_{i=1}^{n} \frac{1}{\phi_i} \mathbb{E}\left[b''(\theta_i)\right]$$
$$\Rightarrow \mathrm{var}\left[y_i\right] \equiv \mathbb{E}\left[(y_i - \mu_i)^2\right] = \phi_i b''(\theta_i). \blacksquare$$

Let us point out that the exponential family distributions are completely characterized by a relation between the mean and variance. In particular, from the formulas (23) and (24) it derives that

$$\mathrm{var}\left[y_i\right] = \phi_i \frac{\partial}{\partial \theta_i} \mu_i. \tag{28}$$

As examples we show below that Normal and Bernoulli distributions are special cases of the general exponential family.

Normal distribution:
The Normal distribution has the density

$$f(y_i|\bar{y}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \bar{y})^2}{2\sigma^2}\right). \tag{29}$$

where $\bar{y}$ is the center of the distribution and σ is the standard deviation. Expanding the square and raising the normalization factor in the exponent we bring (29) in the exponential family form (12) as

$$f(y_i|\bar{y}, \sigma^2) = \exp\left(\frac{y_i \bar{y} - \frac{1}{2}\bar{y}^2}{\sigma^2} - \frac{{y_i}^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right), \tag{30}$$

where we read

$$\theta_i = \bar{y} \quad \text{and} \quad b(\theta_i) = {\theta_i}^2/2, \tag{31}$$

$\phi_i = \sigma^2$, and $c(y_i, \phi_i) = -(2{y_i}^2 + \phi_i \log(2\pi\phi_i))/4\phi_i$. By using the formulas (23) and (24) we find that $\mathbb{E}[y_i] = \theta = \bar{y}$ and $\mathrm{var}[y_i] = \sigma^2$, which are the correct moments for the Normal distribution and agree with the relation (28). The result $\mathbb{E}[y_i] = \theta_i$, that is, the expectation value is proportional to the canonical parameter, reflects the identity link function of the linear regression model which naturally derives from the Normal distribution; this will be discussed extensively in the next section.

Bernoulli distribution:

Bernoulli distribution is the discrete probability distribution of a random variable that takes the values 0 and 1 and has the density

$$f(y_i|p) = p^{y_i}(1-p)^{1-y_i}, \tag{32}$$

which is written, after some algebra, as

$$f(y_i|p) = \exp\left(y_i \log\frac{p}{1-p} + \log(1-p)\right), \tag{33}$$

where we read for the canonical parameter

$$\theta_i = \log\frac{p}{1-p}. \tag{34}$$

Solving for p we obtain $p = e^{\theta_i}/(1+e^{\theta_i})$ and substituting in Eq. (33) yields

$$f(y_i|\theta_i) = \exp\left(y_i\theta_i - \log(1+e^{\theta_i})\right), \tag{35}$$

where we read for the cumulant function

$$b(\theta_i) = \log(1+e^{\theta_i}), \tag{36}$$

as well as, $\phi_i = 1$ and $c(y_i, \phi_i) = 0$. Using the equations (23) and (24) we recover the well known mean and variance for the Bernoulli distribution namely, $\mathbb{E}[y_i] = e^{\theta_i}/(1+e^{\theta_i}) = p$, and $\text{var}[y_i] = e^{\theta_i}/(1+e^{\theta_i})^2 = p(1-p)$, respectively. Bernoulli distribution is a special case of the binomial distribution where a single experiment/trial is conducted. Binomial distribution belongs to the general family of exponential distributions as well. We observe that in Bernoulli distribution the expectation value is not proportional to the canonical parameter θ_i. Subsequently, the linear relation between the expectation value and the predictors is broken. In order to recover the linear relationship we need to generalize the model by introducing the *Link* function which transforms the expectation value to depends linearly on the canonical parameter θ_i and, in turn, on the predictors $\mathbf{x}_i$. This is the second step for the formulation of GLMs and discussed below.

2.3 Link function

In this section we develop the second generalization step for the GMLs formulation. We introduce the *link function* $g(\mu_i)$ that transforms the expectation values to be linear with the predictors. Specifically, we introduce a one-to-one continuous differentiable transformation $g(.)$ and require

$$\eta_i = g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}, \tag{37}$$

where η_i is called the *linear predictor*. We point out that we do not transform the observation data y_i but its expectation value μ_i. For instance, a model where $\log y_i$ is linear on $\mathbf{x}_i$ is not the same as a GLM where $\log\mu_i$ is linear to $\mathbf{x}_i$. Examples of link functions that are investigated in this section include the *identity* and *logit* functions that correspond to the

linear and logistic regression models, respectively. Since the link is a one-to-one function, we can invert it to obtain

$$\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}). \tag{38}$$

When the link function makes the linear predictor the same as the canonical parameter ($\eta_i = g(\mu_i) = \theta_i$), we say that we have a *canonical link*. Subsequently, investigating the relation between θ_i and μ_i yields the link function. An additional relationship that is revealed by using the formula (23) for a canonical link is

$$\begin{aligned} g(\mu_i) = \theta_i \quad &\Rightarrow \quad \mu_i = g^{-1}(\theta_i) \Rightarrow \\ b'(\theta_i) = g^{-1}(\theta_i) \quad &\text{or} \quad g(\theta_i) = (b')^{-1}(\theta_i), \end{aligned} \tag{39}$$

which maps the canonical transformation $g(.)$ to the cumulant function b revealing that b has to be an invertible function. Some examples of the natural pairing between the error distribution and the canonical link function are summarized in the table 1.

Distribution: f_{θ_i}	**Mean Function:** $\mu = b'(\theta)$	**Canonical Link:** $\theta = g(\mu)$
Normal	θ	μ
Bernoulli/Binomial	$e^{\theta}/(1+e^{\theta})$	$\log(\mu/(1-\mu))$
Poisson	e^{θ}	$\log \mu$
Gamma	$-1/\theta$	$-1/\mu$
Inverse Gaussian	$(-2\theta)^{-1/2}$	$-1/(2\mu^2)$

Table 1: Natural pairing between distribution of observations and link functions.

As examples, we calculate the canonical links for the Normal and Bernoulli distributions:
Normal distribution & Linear regression:
From Eq. (31) we have $\theta_i = \mu_i$, subsequently

$$\begin{aligned} \theta = g(\mu) = \mu \Rightarrow \\ g = \text{Identity}. \blacksquare \end{aligned}$$

Bernoulli distribution & Logistic regression:
From Eq. (34) we have $\theta_i = \log[\mu_i/(1-\mu_i)]$, thus

$$\begin{aligned} \theta_i = g(\mu_i) = \log \frac{\mu_i}{1-\mu_i} \Rightarrow \\ g = \text{Logit}. \blacksquare \end{aligned}$$

In the last example we recover the very useful *logistic regression* model that is used when the observations come from binary data. In that case we put the logistic regression in a very general and broad family in the context of exponential distributions proving that linear and logistic regression models are formulated in the same overall framework. Working in such a general framework is a great advantage since there are theory and inferential

methods associated with general GLMs that can be applied afterwards in each specific distribution and regression model. For instance, in the next section we discuss about the maximum likelihood estimation in the very general framework of GLMs and recover the *normal equations* for the likelihood maximization for the linear and logistic regression models.

3 Maximum Likelihood Estimation

In the last section we discuss about *reverse engineering*, that is, we have a dataset of observations $\{y_i, \mathbf{x}_i\}$ and aim to find a conditional distribution that describes them. We derive in general regression models, in the context of likelihood, in which the maximizer is not given by a closed-form and hence we need to use optimization algorithms to compute the maximizer (see advanced section 2).

Let $\{y_i, \mathbf{x}_i\}_{i=1}^n$ be n independent random pairs such that the conditional distribution given $f(y_i|\mathbf{x}_i)$ has density in the canonical exponential family (12). Thus, the likelihood is given by:

$$L(y_i|\theta_i) = \prod_{i=1}^{n} \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right) \tag{40}$$

and the log-likelihood reads

$$\ell(y_i|\theta_i) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + \sum_{i=1}^{n} c(y_i, \phi_i), \tag{41}$$

where the last term is the normalization constant and does not play any role in the maximization of ℓ, hence it can be neglected. Our aim is to determine the $\boldsymbol{\beta}$ of a linear predictor $\mathbf{x}_i^T\boldsymbol{\beta}$ that maximizes the log-likelihood (41). We keep the same vector of coefficients $\boldsymbol{\beta}$ (independent on i) for many pairs of $(y_i, \mathbf{x}_i)$, as more the training dataset $(y_i, \mathbf{x}_i)$ as better the estimation of $\boldsymbol{\beta}$. We first describe maximum likelihood estimation for *canonical links* and find the *normal equations* that maximize the likelihood function. That gives us the intuition to work with general links and derive to the *generalized estimating equations* that maximize the likelihood. In general, the solution of the normal and generalized estimating equations does not have a closed-form and hence, it can be computed by iterated numerical methods such as *Fisher score* and *weighted least squares*, or by using regression modeling packages such as the Python library *Scikit-learn*; the discussion of these iteration methods is out of the scope of these notes.

3.1 Maximum Likelihood Estimation for Canonical Links

When g is the canonical link, and hence the canonical parameter is $\theta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, the log-likelihood takes the simple form

$$\ell(y_i|\mathbf{x}_i^T\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n} w_i\left(y_i\mathbf{x}_i^T\boldsymbol{\beta} - b(\mathbf{x}_i^T\boldsymbol{\beta})\right), \tag{42}$$

where w_i shows the weights of the dispersion parameter, i.e. $\phi_i = \phi/w_i$. Inspecting the Eq. (42) and using Eq. (24), we observe that the second derivative of log-likelihood with respect to $\boldsymbol{\beta}$,

$$\frac{\partial^2 \ell}{\partial^2 \boldsymbol{\beta}^2} = \sum_{i=1}^{n} 0 - \frac{w_i b''(\mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i^2}{\phi} = -\frac{1}{\phi^2} \sum_{i=1}^{n} w_i^2 \text{var}\left[y_i\right] \mathbf{x}_i^2 < 0,$$

is strictly concave, since the variance is always positive, and subsequently, ℓ can be maximized. Taking the partial derivative with respect to $\boldsymbol{\beta}$ yields the score function:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{1}{\phi} \sum_{i=1}^{n} w_i \left(y_i - b'\left(\mathbf{x}_i^T \boldsymbol{\beta}\right)\right) \mathbf{x}_i^T \tag{43}$$

According to the Eq. (23), $\mu_i = b'(\theta_i) = b'\left(\mathbf{x}_i^T \boldsymbol{\beta}\right)$. Now, we can require $\partial_{\boldsymbol{\beta}} \ell = 0$ and solve for the maximum likelihood estimators of $\boldsymbol{\beta}$ through the *normal equations*

$$\sum_{i=1}^{n} w_i \left(y_i - \mu_i\right) \mathbf{x}_i^T = 0. \tag{44}$$

We are seeking for the μ_i that satisfies the above equations. Remember that for canonical links $\mu_i = g^{-1}(\theta_i) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$, hence by solving the equations (44) we essentially estimates the vector of the coefficients $\boldsymbol{\beta}$, which typically has a unique solution.

For example, in the *Normal distribution* the canonical link is $\mu_i = \mathbf{x}_i^T \beta$ (see table table 1). Thus, the formula (44) yields the *linear regression* model

$$\sum_{i=1}^{n} w_i \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right) \mathbf{x}_i^T = 0.$$

Furthermore, in *Bernoulli distribution*, where the canonical link is the logit and the mean is given in the table (1), the expression (44) yields the *logistic regression* model

$$\sum_{i=1}^{n} w_i \left(y_i - \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}\right) \mathbf{x}_i^T = 0.$$

3.2 Maximum Likelihood Estimation for General Links

Rather than canonical, the link can be any function that transforms the expectation value to be linear to the input $\mathbf{x}_i$. For general links, where $\mu_i = b'(\theta_i)$ and $g(\mu_i) = \mathbf{x}_i^T \beta$ (but $g(\mu_i) \neq \theta_i$), the score function reads

$$\begin{aligned}
\frac{\partial}{\partial \beta_j} \ell(y_i|\theta_i) &= \sum_{i=1}^{n} \frac{1}{\phi_i} \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - \frac{\partial b(\theta_i)}{\partial \beta_j}\right) \\
&= \sum_{i=1}^{n} \frac{1}{\phi_i} \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - \frac{\partial b(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j}\right) \\
&= \sum_{i=1}^{n} \frac{1}{\phi_i} \frac{\partial \theta_i}{\partial \beta_j} \left(y_i - \mu_i\right).
\end{aligned} \tag{45}$$

Using the chain rule along with the relations $\mu_i = b'(\theta_i)$ and $\text{var}[y_i] = \phi_i b''(\theta_i)$, we get

$$\begin{aligned}\frac{\partial \mu_i}{\partial \beta_j} &= \frac{\partial b'(\theta_i)}{\partial \beta_j} = \frac{\partial b'(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} \\ &= b''(\theta_i)\frac{\partial \theta_i}{\partial \beta_j} = \frac{\text{var}[y_i]}{\phi_i}\frac{\partial \theta_i}{\partial \beta_j},\end{aligned}$$

hence,

$$\frac{1}{\phi_i}\frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\text{var}[y_i]}\frac{\partial \mu_i}{\partial \beta_j}. \tag{46}$$

Substituting the relation (46) into the score function (45) and maximizing (request Eq. (45) to be zero), we obtain the *generalized estimating equations*

$$\sum_{i=1}^{n} \frac{1}{\text{var}[y_i]}\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}(y_i - \mu_i) = 0, \tag{47}$$

where the roots imply the maximum likelihood estimates. It is easy to confirm that in the case of a canonical link where $\mu_i = b'(\theta_i) = b'(\mathbf{x}_i^T\beta)$, the generalized estimating equations (47) yield the normal equations (44).

References

[1] C. Bishop, *Pattern Recognition and Machine Learning*, 8th ed. Springer (2008).

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 8th ed., Springer (2017).

[3] A. Agresti, *Foundations of Linear and Generalized Linear Models*, Wiley (2015).

[4] J. A. Nelder and R. W. M. Wedderburn, *Generalized Linear Models*, J. of the Royal Statistical Society. Series A (General), **135**, No. 3, 370-384 (1972). https://www.jstor.org/stable/2344614