Generalized Linear Models (GLM)

Xiaogang Su, Ph.D.
Department of Mathematical Sciences
University of Texas at El Paso (UTEP)
xsu@utep.edu

Contents

Linear regression has wide and fundamental applications in various fields. Its popularity can be attributed to its simple form, sound theoretical support, efficient computation in estimation, great flexibility to incorporate interactions, dummy variables, and other transformations, and easy interpretations. For all the linear models discussed so far, the response Y is a continuous variable. Many studies or experiments are often involve responses of other types. Consider, for example, evaluation of the academic performance of college students. Instead of the continuous grade point average (GPA) score, the categorical grade A–F may have been used. Generalized linear models (GLM; McCullagh and Nelder, 1983) extends linear regression to encompass other types of response while, at the same time, enjoying nearly all the merits of linear modeling.

In this chapter, we study how linear regression is generalized to handle data with different types of responses. We first motivate the problem using an example on simple logistic regression in Section 1, followed by general discussion on the basic components (Section 2), estimation (Section 3), statistical inference, and other issues (Section 4) in GLM. We then introduce two important and very commonly used GLMs, logistic regression models for binary responses in Section 5 and log-linear models for count data in Section 6.

1 Introduction: A Motivating Example

We first reproduce an example from Hosmer and Lemeshow (2000) that gives an excellent motivation to the problem with a real application. The data set, which can be downloaded at

`ftp://ftp.wiley.com/public/sci_tech_med/logistic/alr.zip`,

was collected from a retrospective coronary heart disease (CHD) study. It contains three variables: ID for each subject, AGE (X), and a binary indicator CHD (Y) indicating whether CHD occurs to the subject. The objective is to explore the relationship between age and prevalence of CHD.

A scatterplot of the data is given in Fig. 1(a). Although one can see larger values of AGE tend to be more associated with "1"'s of CHD, the plot is not very informative due to discreteness of the response Y.

Recall that linear regression relates the conditional mean of the response, $E(Y|X)$, to a linear combination of predictors. Can we do the same with binary data? Let the binary response $y_i = 1$ if CHD is found in the i-th individual and 0 otherwise for $i = 1, \ldots, n$. Denote $\pi_i = \Pr(y_i = 1)$. Thus, the conditional distribution of y_i given x_i is a Bernoulli trial with parameter π_i. It can be found that $E(y_i) = \pi_i$. Hence, the linear model would be

$$E(y_i) = \pi_i = \beta_0 + \beta_1 x_i. \tag{1}$$

Fig. 1(a) plots the straight line fitted by least squares. Clearly, it is not a good fit. There is another inherent problem with Model (1). The left-hand side π_i ranges from 0 to 1, which does not mathematically match well with the range $(-\infty, \infty)$ of the linear equation on the right-hand side. A transformation on π_i, $g(\cdot)$, which maps $[0, 1]$ onto $(-\infty, \infty)$, would help. This transformation function is referred to the link function.

In order to explore the functional form between π_i and x_i, we must have available estimates of the proportions π_i. One approach is group the data by categorizing AGE into several intervals and record the relatively frequency of CHD within each interval. Table 1 shows the worksheet for this calculation.

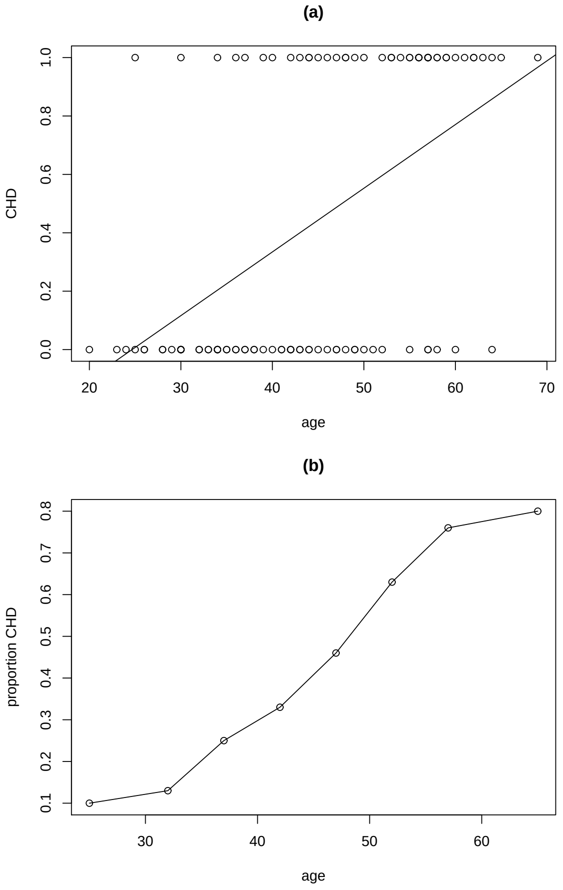

Figure 1: (a) Scatterplot of the CHD data, superimposed with the straight line from least squares fit; (b) Plot of the percentage of subjects with CHD within each age group, superimposed by LOWESS smoothed curve.

Fig. 1(b) plots the proportions of subjects with CHD in each age interval versus the middle value of the interval. It can be seen that the conditional mean of y_i or proportion gradually approaches zero and one to each end. The plot shows an 'S'-shaped or sigmoid nonlinear relationship, in which the change in $\pi(x)$ with per unit increase in x grows quickly at first, but gradually slows down and then eventually levels off. Such a pattern is rather representative and can be generally seen in many other applications. It is often expected to see that a fixed change in x has less impact when $\pi(x)$ is near 0 or 1 than when $\pi(x)$ is near 0.5. Suppose, for example, that $\pi(x)$ denotes the probability to pass away for a person of age x. An increase of five years in age would have less effect on $\pi(x)$ when $x = 70$, in which case $\pi(x)$ is perhaps close to 1, than when $x = 40$.

In sum, a suitable link function $g(\pi_i)$ is desired to satisfy two conditions: it maps $[0, 1]$ onto the whole real line and has the sigmoid shape. A natural choice for $g(\cdot)$ would be a cumulative distribution function of a random variable. In particular, the logistic distribution, whose CDF is the simplified logistic function $g(x) = \exp(x)/\{1 + \exp(x)\}$, yields the most popular link. Under the logistic link, the relationship between the CHD prevalence rate and AGE can be formulated by

Table 1: Frequency Table of AGE Group by CHD.

		CHD		
Age Group	n	Absent	Present	Proportion
20-29	10	9	1	0.10
30-34	15	23	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
Total	100	57	43	0.43

the following simple model

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i.$$

When several predictors $\{X_1, \ldots, X_p\}$ are involved, the multiple logistic regression can be generally expressed as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}.$$

We shall explore more on logistic regression in Section 5.

2 Components of GLM

The logistic regression model is one of the generalized linear models (GLM). Many models in the class had been well studied by the time when McCullagh and Nelder (1983) introduced the unified GLM family. The specification of a GLM generally consists of three components: a *random component* specifies the probability distribution of the response; a *systematic component* forms the linear combination of predictors; and a link function relates the mean response to the systematic component.

2.1 Exponential Family

The random component assumes a probability distribution for the response y_i. This distribution is taken from the natural exponential distribution family of form

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi)\right\}, \tag{2}$$

where θ_i is the *natural parameter* and ϕ is termed as the dispersion parameter. It can be shown that

$$E(y_i) = \mu_i = b'(\theta_i) \quad \text{and} \quad \text{var}(y_i) = b''(\theta_i)a(\phi), \tag{3}$$

both moments determined by the function $b(\cdot)$.

Take the Gaussian distribution for example. The probability density function (pdf) of $N(\mu, \sigma^2)$ can be rewritten as

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

Therefore, $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$, and $c(y, \phi) = -\left\{y^2/\phi + \log(2\pi\phi)\right\}/2$.

2.2 Linear Predictor and Link Functions

The systematic component is the *linear predictor*, denoted as

$$\eta_i = \sum_j \beta_j x_{ij} = \mathbf{x}_i \boldsymbol{\beta},$$

for $i = 1, \ldots, n$ and $j = 0, 1, \ldots, p$ with $x_{i0} = 1$ to account for the intercept. Similar to ordinary linear regression, this specification allows for incorporation of interaction, polynomial terms, and dummy variables.

The *link function* in GLM relates the linear predictor η_i to the mean response μ_i. Thus

$$g(\mu_i) = \eta_i$$

or inversely

$$\mu_i = g^{-1}(\eta_i).$$

In classical Gaussian linear models, the identity link $g(\mu_i) = \mu_i$ is applied. A preferable link function usually not only maps the range of μ_i onto the whole real line, but also provides good empirical approximation and carries meaningful interpretation when it comes to real applications.

As an important special case, the link function g such that

$$g(\mu_i) = \theta_i$$

is called the *canonical link*. Under this link, the direct relationship $\theta_i = \eta_i$ occurs. Since $\mu_i = b'(\theta_i)$, we have $\theta_i = (b')^{-1}(\mu_i)$. Namely, the canonical link is the inverse of $b'(\cdot)$:

$$(b')^{-1}(\mu_i) = \eta_i = \sum_j \beta_j x_{ij}. \tag{4}$$

In Gaussian linear models, the canonical link is the identity function. With the canonical link, the sufficient statistic is $\mathbf{X}^T\mathbf{y}$ in vector notation with components $\sum_i x_{ij} y_i$ for $j = 0, 1, \ldots, p$. The canonical link provides mathematical convenience in deriving statistical properties of the model; at the same time, they are also often found eminently sensible on scientific grounds.

3 Maximum Likelihood Estimation of GLM

The least squares (LS) method is no longer directly appropriate when the response variable Y is not continuous. Estimation of GLM is processed within the maximum likelihood (ML) framework. However, as we will see, the ML estimation in GLM has a close connection with an iteratively weighted least squares method.

Given data $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$, the log likelihood function is

$$L(\boldsymbol{\beta}) = \sum_i L_i = \sum_i \log f_Y(y_i; \theta_i, \phi) = \sum_i \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_i c(y_i, \phi). \tag{5}$$

3.1 Likelihood Equations

The likelihood equations are

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_i \frac{\partial L_i}{\partial \beta_j} = 0$$

for $j = 0, 1, \ldots, p$. Using the chain rule, we have

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j},$$

where

$$\begin{aligned}
\frac{\partial L_i}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)} \quad \text{using} \quad \mu_i = b'(\theta_i); \\
\frac{\partial \theta_i}{\partial \mu_i} &= b''(\theta_i) = \text{var}(y_i)/a(\phi) \quad \text{using} \quad \text{var}(y_i) = b''(\theta_i)a(\phi); \\
\frac{\partial \mu_i}{\partial \eta_i} &= (g^{-1})'(\eta_i); \\
\text{and } \frac{\partial \eta_i}{\partial \beta_j} &= x_{ij}.
\end{aligned}$$

Therefore, the likelihood equations for $\boldsymbol{\beta}$ become

$$\sum_{i=1}^n \frac{(y_i - \mu_i)\, x_{ij}}{\text{var}(y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad \text{for } j = 0, 1, \ldots, p. \tag{6}$$

In the case of a canonical link $\eta_i = \theta_i$, we have

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i).$$

Thus

$$\frac{\partial L_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\text{var}(y_i)} b''(\theta_i) x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{a(\phi)} \tag{7}$$

using (3) and the likelihood equations simplify to

$$\sum_i x_{ij} y_i = \sum_i x_{ij} \mu_i, \text{ for } j = 0, 1, \ldots, p.$$

Or in matrix notations, $\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{0}$, which is in the same form as seen in ordinary linear regression

$$\mathbf{X^T X}\boldsymbol{\beta} = \mathbf{X^T y} \implies \mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{0}.$$

3.2 Fisher's Information Matrix

Fisher's information matrix $\mathcal{I}$ is defined as the negative expectation of the second derivatives of the log-likelihood with respect to $\boldsymbol{\beta}$, i.e., $\mathcal{I} = E(-L'')$ with elements $E\{-\partial^2 L(\boldsymbol{\beta})/\partial\beta_j\partial\beta_{j'}\}$.

Using the general likelihood results

$$E\left(\frac{\partial^2 L_i}{\partial\beta_j\partial\beta_{j'}}\right) = -E\left(\frac{\partial L_i}{\partial\beta_j}\frac{\partial L_i}{\partial\beta_{j'}}\right),$$

which holds for distributions in the exponential family (Cox and Hinkley,, 1974, Sec. 4.8), we have

$$\begin{aligned} E\left(\frac{\partial L_i}{\partial\beta_j}\frac{\partial L_i}{\partial\beta_{j'}}\right) &= -E\left\{\frac{(y_i-\mu_i)x_{ij}}{\mathrm{var}(y_i)}\frac{\partial\mu_i}{\partial\eta_i}\frac{(y_i-\mu_i)x_{ij'}}{\mathrm{var}(y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right\} \quad \text{from (6)} \\ &= -\frac{x_{ij}x_{ij'}}{\mathrm{var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \end{aligned}$$

and hence

$$E\left(-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\beta_{j'}}\right) = \sum_{i=1}^n \frac{x_{ij}x_{ij'}}{\mathrm{var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \tag{8}$$

Let $\mathbf{W} = \mathrm{diag}(w_i)$ be the diagonal matrix with diagonal elements

$$w_i = \frac{1}{\mathrm{var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \tag{9}$$

Then the information matrix is given by

$$\mathcal{I} = \mathbf{X^T W X} \tag{10}$$

If the canonical link is used, then from equation (7)

$$\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\beta_{j'}}\right) = -\frac{x_{ij}}{a(\phi)}\cdot\frac{\partial\mu_i}{\partial\beta_{j'}},$$

which does not involve the random variable y_i. This implies that

$$\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\beta_{j'}}\right) = E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\beta_{j'}}\right). \tag{11}$$

In other words, the *observed information* matrix is equal to the *expected information* matrix with the canonical link.

3.3 Optimization of the Likelihood

The log-likelihood function of a GLM is typically nonlinear in $\boldsymbol{\beta}$. Optimization of the likelihood is usually done via iterative numerical algorithms such as the *Newton-Raphson method* or Fisher scoring method.

Denote the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}$, which satisfies $L'(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$. Applying the first-order Taylor series expansion on $L'(\widehat{\boldsymbol{\beta}})$ at the current estimate $\boldsymbol{\beta}^{(k)}$ gives

$$0 = L'(\widehat{\boldsymbol{\beta}}) \approx L'(\boldsymbol{\beta}^{(k)}) + L''(\boldsymbol{\beta}^{(k)})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(k)}).$$

Solving the equation for $\widehat{\boldsymbol{\beta}}$ leads to the Newton-Raphson updating formula

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \left(\mathbf{H}^{(k)}\right)^{-1} \mathbf{u}^{(k)}, \tag{12}$$

where $\mathbf{u}^{(k)} = L'(\boldsymbol{\beta}^{(k)})$ is the first derivative or *gradient* of the log-likelihood evaluated at $\boldsymbol{\beta}^{(k)}$ and $\mathbf{H}^{(k)} = L''(\boldsymbol{\beta}^{(k)})$ is its second derivative or *Hessian matrix* evaluated at $\boldsymbol{\beta}^{(k)}$.

Fisher scoring resembles the Newton-Raphson method, except for that Fisher scoring uses the expected value of $-\mathbf{H}^{(k)}$ (i.e., the expected information), whereas Newton-Raphson applies the observed information directly. Note that $\mathcal{I}^{(k)} = E\left(-\mathbf{H}^{(k)}\right)$. Plugging it into expression (12) yields the updating formula for Fisher scoring

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \left(\mathcal{I}^{(k)}\right)^{-1} \mathbf{u}^{(k)}. \tag{13}$$

It is worth noting that, from (11), the Newton-Raphson method is identical to Fisher scoring under the canonical link, in which case the observed information is non-random and hence equal to the expected information.

Next, we shall show that implementation of Fisher scoring in GLM takes the form of an *iteratively reweighted least squares* algorithm. The weighted least squares (WLS) estimator of $\boldsymbol{\beta}$ is referred to

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{z}, \tag{14}$$

when the linear model is

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \mathbf{V})$.

From (6), the component of the gradient $\mathbf{u}$ is

$$u_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\, x_{ij}}{\mathrm{var}(y_i)} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \cdot \frac{\partial \eta_i}{\partial \mu_i}.$$

Hence, $\mathbf{u}$ can be rewritten in matrix form as

$$\mathbf{u} = \mathbf{X}^T\mathbf{W}\boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu}), \tag{15}$$

where $\boldsymbol{\Delta} = \mathrm{diag}\left(\partial\eta_i/\partial\mu_i\right)$ and $\mathbf{W}$ is given in equation (9).

Also, from Equation (10), $\mathcal{I} = \mathbf{X}^T\mathbf{W}\mathbf{X}$. Therefore,

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} + \left(\mathcal{I}^{(k)}\right)^{-1}\mathbf{u}^{(k)}, \text{ where } \mathcal{I}^{(k)} = \mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X} \text{ from (10)}, \\
&= (\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X}\boldsymbol{\beta}^{(k)} + \\
&\quad (\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(k)}\boldsymbol{\Delta}^{(k)}(\mathbf{y} - \boldsymbol{\mu}^{(k)}) \\
&= \left\{\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X}\right\}^{-1}\mathbf{X}^T\mathbf{W}^{(k)}\left\{\mathbf{X}\boldsymbol{\beta}^{(k)} + \boldsymbol{\Delta}^{(k)}(\mathbf{y} - \boldsymbol{\mu}^{(k)})\right\} \\
&= \left\{\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X}\right\}^{-1}\mathbf{X}^T\mathbf{W}^{(k)}\left\{\boldsymbol{\eta}^{(k)} + \boldsymbol{\Delta}^{(k)}(\mathbf{y} - \boldsymbol{\mu}^{(k)})\right\} \\
&= \left\{\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X}\right\}^{-1}\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{z}^{(k)} \qquad (16)
\end{aligned}$$

if one defines an adjusted response

$$\mathbf{z}^{(k)} = \boldsymbol{\eta}^{(k)} + \boldsymbol{\Delta}^{(k)}(\mathbf{y} - \boldsymbol{\mu}^{(k)})$$

with components

$$z_i^{(k)} = \eta_i^{(k)} + \left(y_i - \mu_i^{(k)}\right) \cdot \left.\frac{\partial \eta_i}{\partial \mu_i}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}} \qquad (17)$$

for $i = 1, \ldots, n$. Note that $z_i^{(k)}$'s are continuously scaled. Comparing (16) to (14), it is clear that $\boldsymbol{\beta}^{(k+1)}$ is the WLS solution for fitting ordinary linear model

$$\mathbf{z}^{(k)} = \mathbf{X}\boldsymbol{\beta}^{(k)} + \boldsymbol{\varepsilon}, \quad \text{with } \boldsymbol{\varepsilon} \sim \left\{\mathbf{0}, \left(\mathbf{W}^{(\mathbf{k})}\right)^{-1}\right\}.$$

4 Statistical Inference and Other Issues in GLM

4.1 Wald, Likelihood Ratio, and Score Test

The maximum likelihood framework is a well-established system for statistical inference. The maximum likelihood estimators have many attractive properties. For example, they are asymptotically consistent and efficient. Large-sample normality is also readily available under weak regularity conditions. Standard ML results supplies that

$$\widehat{\boldsymbol{\beta}} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\beta}, \ \mathcal{I}^{-1}\right) \quad \text{or} \quad \mathcal{N}\left\{\boldsymbol{\beta}, \ \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\right\} \text{ as } n \to \infty \qquad (18)$$

where the notation $\xrightarrow{d}$ means "converges in distribution to." The asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\text{cov}(\widehat{\boldsymbol{\beta}}) = \mathcal{I}^{-1} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1},$$

which can be estimated by replacing $\mathbf{W}$ with its estimate $\hat{\mathbf{W}}$. Another immediate ML result is the so-called delta method, which can be used to obtain asymptotic distributions for any smooth function of $\widehat{\boldsymbol{\beta}}$, $g(\widehat{\boldsymbol{\beta}})$. Let $\mathbf{g}' = \partial g(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. Then

$$g(\widehat{\boldsymbol{\beta}}) \xrightarrow{d} \mathcal{N}\left\{g(\boldsymbol{\beta}), \ (\mathbf{g}')^T\mathcal{I}^{-1}\mathbf{g}'\right\}, \qquad (19)$$

where the asymptotic variance $\text{var}\{g(\widehat{\boldsymbol{\beta}})\} = \mathbf{g}'^T \mathcal{I}^{-1} \mathbf{g}'$ can be estimated by substituting $\boldsymbol{\beta}$ with its MLE $\widehat{\boldsymbol{\beta}}$. This result can be heuristically justified by the Taylor expansion

$$g(\widehat{\boldsymbol{\beta}}) - g(\boldsymbol{\beta}) \approx (\mathbf{g}')^T (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Within the ML framework, there are three commonly-used methods for making statistical inference on $\boldsymbol{\beta}$: the Wald test, the likelihood ratio test, and the score test, all exploiting the asymptotic normality of maximum likelihood estimation. We will briefly discuss each of them from the hypothesis testing prospective. Confidence intervals can be generally derived by inverting the testing procedures.

The Wald statistic for testing the null H_0: $\boldsymbol{\Lambda}\boldsymbol{\beta} = \mathbf{b}$ where $\boldsymbol{\Lambda}$ is $q \times (p+1)$ of rank q takes a similar form seen in ordinary linear regression:

$$G = \left(\boldsymbol{\Lambda}\widehat{\boldsymbol{\beta}} - \mathbf{b}\right)^T \left\{\boldsymbol{\Lambda}\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\boldsymbol{\Lambda}^T\right\}^{-1} \left(\boldsymbol{\Lambda}\widehat{\boldsymbol{\beta}} - \mathbf{b}\right). \tag{20}$$

The null distribution of G is referred to $\chi^2(q)$. As a special case, for H_0: $\beta_j = b$, the z test

$$z = \sqrt{G} = \frac{\hat{\beta}_j - b}{SE(\hat{\beta}_j)} \overset{H_0}{\sim} N(0,1)$$

can be used. It is most convenient to derive confidence intervals from the Wald test. For example, $(1-\alpha) \times 100\%$ CI for β_j is given by

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}}\, SE(\hat{\beta}_j).$$

Clearly, Wald-typed inference is also readily available for functions of $\boldsymbol{\beta}$, $g(\boldsymbol{\beta})$, using results in (19).

The likelihood ratio test (LRT) compares the maximized log-likelihood functions between two nested models. Suppose that we have a model, called the *full model*, and a null hypothesis H_0. Bringing the conditions in H_0 into the full model gives a *reduced* model. Let $\widehat{L}_{\text{full}}$ and $\widehat{L}_{\text{reduced}}$ denote their respective maximized log-likelihoods. The likelihood ratio test statistic is

$$LRT = -2 \cdot \left(\widehat{L}_{\text{reduced}} - \widehat{L}_{\text{full}}\right). \tag{21}$$

The null distribution of LRT is again $\chi^2(\nu)$, where ν is the difference in number of degrees of freedom between two models.

The third alternative for testing $H_0 : \boldsymbol{\beta} = \mathbf{b}$ is the score test, which is also the Lagrange multiplier test. It is based on the slope and expected curvature of $L(\boldsymbol{\beta})$ at the null value $\mathbf{b}$. Let

$$\mathbf{u}_0 = \mathbf{u}(\mathbf{b}) = \left.\frac{\partial L}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}\,=\,\mathbf{b}}$$

be the score function of the full model evaluated at the null value $\mathbf{b}$ (recall that $\mathbf{u}(\hat{\boldsymbol{\beta}}) = 0$) and

$$\mathcal{I}_0 = \mathcal{I}(\mathbf{b}) = -\,E\left(\frac{\partial^2 L}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}^T}\right)\Bigg|_{\boldsymbol{\beta}\,=\,\mathbf{b}}$$

be the Fisher's information matrix of the full model evaluated at $\mathbf{b}$. The score test statistic is a quadratic form given as

$$S = \mathbf{u}_0^T \, \mathcal{I}_0^{-1} \, \mathbf{u}_0. \tag{22}$$

Under H_0, the score test statistic follows the same chi-squared null distribution. Consider another illustration, which is more practically useful. Suppose that $\boldsymbol{\beta}$ can be partitioned into $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$ and we want to test $H_0 : \boldsymbol{\beta}_1 = \mathbf{b}_1$. Let $\hat{\boldsymbol{\beta}}_2^{(0)}$ denote the MLE of $\boldsymbol{\beta}_2$ obtained from fitting the reduced model or the null model. The score test statistic is then given by

$$S = \mathbf{u}_1^T \, \left(\mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21}\right)^{-1} \, \mathbf{u}_1 \Big|_{\boldsymbol{\beta} = \, (\mathbf{b}_1, \, \hat{\boldsymbol{\beta}}_2^{(0)})^T}, \tag{23}$$

where

$$\mathbf{u}_1 = \frac{\partial L}{\partial \boldsymbol{\beta}_1} \text{ and } \mathbf{I}_{jj'} = \frac{\partial^2 L}{\partial \boldsymbol{\beta}_j \, \partial \boldsymbol{\beta}_{j'}^T}$$

for $j, j' = 1, 2$. Note that all the quantities involved in the above expression are derived from the full model but evaluated at $\boldsymbol{\beta} = (\mathbf{b}_1, \, \hat{\boldsymbol{\beta}}_2^{(0)})^T$, which are obtained from fitting the null model. The null distribution of S is χ^2 with df equal to the dimension of $\boldsymbol{\beta}_1$.

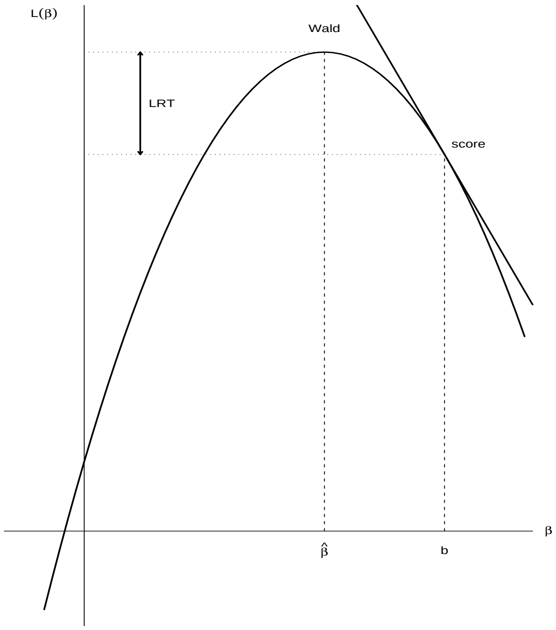

Figure 2: Plot of the log-likelihood function: comparison of Wald, LRT, and score tests on H_0 : $\beta = b$.

The Wald, LRT, and score tests all refer to the same null chi-squared distribution and they can be shown to be asymptotically equivalent for large sample sizes. On the other hand, they show different empirical performances. Fig. 2 illustrated the comparison of these three tests for testing $H_0 : \beta = b$ in the one-dimensional setting $L(\beta)$. The Wald test utilizes estimation of the full model. Its form is very similar to what we have in ordinary linear regression and hence easy to comprehend. For this reason, confidence intervals are often derived from Wald tests. The score test is solely based on the null or reduced model estimation. A heuristic derivation of the score test can be carried out as follows. At the null point b, the tangent line of the loglikelihood function $L(\cdot)$ is given by

$$y - L(b) = u(b)(x - b) \tag{24}$$

with slope

$$u(b) = \left.\frac{\partial L}{\partial \beta}\right|_{\beta=b}$$

is the score function or the first derivative of $L(\cdot)$, evaluated at b. Note that $u(\hat{\beta}) = 0$ as the tangent line at $\hat{\beta}$ is flat. Thus how different $u(b)$ is from $u(\hat{\beta}) = 0$ naturally signalizes how different $\hat{\beta}$ is from b. Furthermore, applying Taylor expansion

$$0 = u(\hat{\beta}) \approx u(b) + \mathcal{I}(b)(\hat{\beta} - b)$$

yields $u(b) = -\mathcal{I}(b)(\hat{\beta} - b)$. Under $H_0 : \beta = b$, $\hat{\beta} \xrightarrow{d} \mathcal{N}\{b, \mathcal{I}^{-1}(b)\}$. It follows that, under H_0, $u(b) \xrightarrow{d} \mathcal{N}\{0, \mathcal{I}(b)\}$. Therefore, the score test is $S = u(b)^2/\mathcal{I}(b) \overset{H_0}{\sim} \chi^2(1)$. The LRT combines information from both models and is the more resourceful. In fact, the LRT can be shown to be the most powerful test asymptotically. Often more complex in its specific form, the score test is advantageous computationally as its calculation only requires estimation of the null model, i.e., the reduced model under the null hypothesis. This property renders the score test a very attractive technique in many scenarios where computational efficiency is a major concern. Examples include evaluation of the added variables in stepwise selection and evaluation of allowable splits in recursive partitioning.

4.2 Model Selection

The model selection techniques in linear regression can be extended in its entirety into GLM as well. The AIC criterion, for example, is

$$AIC = -2\,\hat{L} + 2 \times \text{number of parameters}$$

and

$$BIC = -2\,\hat{L} + \log(n) \times \text{number of parameters},$$

where $\hat{L}$ is the maximized log-likelihood.

In terms of L_1 regularization, there are two main approaches available. One is via quadratic approximation (Wang and Leng; *JASA*, 2007) by observing that

$$\begin{aligned} L(\boldsymbol{\beta}) &\approx L(\hat{\boldsymbol{\beta}}) + \dot{L}(\boldsymbol{\beta})^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \ddot{L}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ &\approx L(\hat{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \ddot{L}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ when ALL predictors are included in the model (thus $\dot{L}(\hat{\boldsymbol{\beta}}) = 0$). Recall that the inverse of the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\hat{\Sigma}^{-1} = -\ddot{L}(\hat{\boldsymbol{\beta}})/n$. Therefore, maximizing $L(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is equivalent to minimizing

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \hat{\Sigma}^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}),$$

which becomes a least square problem. With L_1 regularization, the objective function in adaptive lasso is

$$Q(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \hat{\Sigma}^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \sum_{j=1}^{p} \lambda_j |\beta_j|.$$

The efficient LAR algorithm can be readily used for the minimization. The shortcoming of this method is that $\hat{\boldsymbol{\beta}}$ from the full model might not be available in many scenarios such as $p \gg n$ problems.

The second approach is the coordinate descent method (Friedman et al., 2007) implemented in the `glmnet` package in R. This general method works for lasso, and many other related methods (elastic net, grouped lasso). It is an extremely simple-minded approach. The main idea is to minimize over one parameter at a time, keeping all others fixed. Coordinate descent algorithm applies soft-thresholding of a partial residual, similar to the backfitting algorithm in GAM. The essential step involved in coordinate descent is illustrated below, with ordinary multiple linear regression:

$$\hat{\beta}_j(\lambda) \longleftarrow S\left\{ \sum_{i=1}^{n} x_{ij}(y_i - \hat{y}_i^{(j)}),\, \lambda \right\},$$

where the function $S(t, \lambda) = \text{sign}(t)(|t| - \lambda)_+$ provides the soft threshold and $\hat{y}_i^{(j)} = \sum_{k \neq j} x_{ik} \hat{\beta}_k(\lambda)$. Note that the term $(y_i - \hat{y}_i^{(j)})$ is exactly the partial residual for x_j. The algorithm start with a large value of λ. Run procedure until convergence. Then decrease λ using previous solution as a warm start. The key point of CD is that quantities in the above equation can be quickly updated a $j = 1, 2, \ldots, p; 1, 2, \ldots, p; 1 \ldots.$

4.3 Other Model Fitting Issues

Many other methods and procedures of linear regression are readily extended to generalized linear models. In the following, we shall briefly discuss several important aspects. First, the sum of squares error (SSE), also named as residual sum of squares (RSS), is an important lack-of-fit measure in linear regression. In GLM, the *deviance* plays the same role as SSE. Let $\hat{L}$ denote the maximized log likelihood score for the current model and $\hat{L}_{\text{max}}$ denote the maximum possible log likelihood score for the given data, achieved by the so-called *saturated model.* The deviance is defined as

$$D = 2 \times \left(\hat{L}_{\text{max}} - \hat{L} \right).$$

For nested models, a larger deviance is always associated with a reduced or simpler model. The *analysis of deviance* compares two nested models with an LRT chi-squared test, analogous to the use of analysis of variance (ANOVA) in linear regression. Similar to linear regression, categorical predictors are handled by defining dummy variables. Interaction among variables are often quantified by cross-product terms.

For model diagnostic purposes, two types of residuals are commonly used in GLM. The first type uses components of the deviance contributed by individual observations. Let $D = \sum_{i=1}^{n} d_i$. The *deviance residual* for i-th observation is $\sqrt{d_i} \cdot \text{sign}(y_i - \hat{\mu}_i)$. An alternative is the *Pearson residuals*, defined as

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\text{var}}(y_i)}}. \tag{25}$$

Besides, the *hat matrix*, whose diagonal elements supply the *leverage*, is given by

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^{1/2}.$$

Various jackknife-based diagnostic measures, such as Cook's distance, also find their natural extensions in GLM.

A detailed description of all GLM fitting aspects is beyond the scope of this book. we refer interested readers to McCullagh and Nelder (1989) for a full account.

5 Logistic Regression for Binary Data

5.1 Interpreting the Logistic Model

A logistic regression model is a GLM for modeling data with binary responses. Consider data that contain n observations $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$, where y_i is the binary 0-1 response for the i-th individual and $\mathbf{x}_i$ is its associated predictor vector.

A natural model for y_i is the Bernoulli trial with parameter $\pi_i = E(y_i) = \mu_i = \mathrm{P}\{y_i = 1\}$. The probability distribution function of y_i can be written in the exponential family form

$$\begin{aligned} f_Y(y_i) &= \pi_i^{y_i}(1-\pi_i)^{1-y_i} = (1-\pi_i)\left\{\frac{\pi_i}{1-\pi_i}\right\}^{y_i} \\ &= \exp\left\{y_i \log\frac{\pi_i}{1-\pi_i} + \log(1-\pi_i)\right\} \\ &= \exp\left\{\frac{y_i\log\{\pi_i/(1-\pi_i)\} - \{-\log(1-\pi_i)\}}{1} + 0\right\} \end{aligned}$$

Correspondingly, we have $\theta_i = \log\{\pi_i/(1-\pi_i)\}$, $a(\phi) = 1$, $c(y_i, \phi) = 0$, and

$$b(\theta_i) = -\log(1-\pi_i) = \log\left(\frac{1}{1-\pi_i}\right) = \log\left(1 + \frac{\pi_i}{1-\pi_i}\right) = \log\left(1 + e^{\theta_i}\right).$$

It follows that $E(y_i|\mathbf{x}_i) = \pi_i$ and $\mathrm{var}(y_i|\mathbf{x}_i) = \pi_i(1-\pi_i)$.

The logistic regression model applies the canonical link $\theta_i = \eta_i$, which leads to the following formulation:

$$\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta}, \tag{26}$$

or, equivalently,

$$\pi_i = \mathrm{logistic}\left(\mathbf{x}_i^T\boldsymbol{\beta}\right) = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})}. \tag{27}$$

Interpretation of the regression coefficients in $\boldsymbol{\beta}$ in logistic regression, extracted analogously as in linear regression, has to do with the so-called *odds ratio.* The quantity

$$\frac{\pi}{1-\pi} = \frac{\mathrm{P}(Y=1|\mathbf{x}_i)}{\mathrm{P}(Y=0|\mathbf{x}_i)}$$

is often referred to as the *odds* of having $Y = 1$ conditioning on $\mathbf{x}_i$, which is a critical risk measure in many applications. The logistic model can be expressed in terms of odds

$$\log(\mathrm{odds}) = \mathbf{x}_i^T\boldsymbol{\beta}. \tag{28}$$

The ratio of two odds, each from a different scenario, is termed as the *odds ratio* (OR). The odds ratio is an appealing measure for comparing risks. For example,

$$OR = \frac{\mathrm{P}(Y=1|X=1)/\mathrm{P}(Y=0|X=1)}{\mathrm{P}(Y=1|X=0)/\mathrm{P}(Y=0|X=0)}$$

is the ratio of odds for having $Y=1$ between two different states: $X=1$ vs. $X=0$. If $Y = 1\{\text{lung cancer is present}\}$ indicates the status of lung cancer for an individual and $X = 1\{\text{smoker}\}$ indicates whether he or she is a smoker, then $OR = 3$ implies that the odds of developing lung cancer for smokers is three times as much as that for non-smokers.

With similar arguments in linear regression, model (28) implies that every one-unit increase in X_j, while holding other predictors fixed, would lead to an amount of β_j change in the logarithm of the odds. That is

$$\beta_j = \log\left(\mathrm{Odds}_{X_j=x+1}\right) - \log\left(\mathrm{Odds}_{X_j=x}\right) = \log\left(OR_{(x+1):x}\right).$$

In other words, the odds ratio comparing $X_j = x+1$ vs. $X_j = x$, with other predictor fixed, is $OR_{(x+1):x} = \exp(\beta_j)$.

5.2 Estimation of the Logistic Model

The log likelihood for the logistic model (26) is

$$L(\boldsymbol{\beta}) = \sum_i^n \left\{ y_i \log \frac{\pi_i}{1-\pi_i} + \log(1-\pi_i) \right\}.$$

The regression coefficients in $\boldsymbol{\beta}$ enter the log likelihood through its relationship with π_i in (26). But it is often a tactical manoeuvre not to make the direct substitution for π_i. Instead, differentiation of the log-likelihood with respect to $\boldsymbol{\beta}$ is done via the chain rule.

It can be found from equations (6) and (7) that the gradient

$$\mathbf{u} = \frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{y} - \boldsymbol{\pi}),$$

where $\boldsymbol{\pi} = (\pi_i)$ is the vector of expected values. From equations (9) and (10), the Fisher's information matrix is

$$\mathcal{I} = \mathbf{X}^T \mathrm{diag}\{\pi_i(1-\pi_i)\}\mathbf{X}.$$

Thus the updating formula in Fisher scoring becomes, according to (13),

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \left\{\mathbf{X}^T \mathrm{diag}\{\pi_i^{(k)}(1-\pi_i^{(k)})\}\mathbf{X}\right\}^{-1} \mathbf{X}^T(\mathbf{y} - \boldsymbol{\pi}^{(k)}).$$

When implemented with the iterative reweighted least squares, the redefined response in (17) at each intermediate step would be

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta}^{(k)} + \mathrm{diag}\left\{ \left(\pi_i^{(k)}(1-\pi_i^{(k)})\right)^{-1} \right\} (\mathbf{y} - \boldsymbol{\pi}^{(k)}),$$

with components

$$z_i^{(k)} = \log \frac{\pi_i^{(k)}}{1-\pi_i^{(k)}} + \frac{y_i - \pi_i^{(k)}}{\pi_i^{(k)}(1-\pi_i^{(k)})}.$$

For massive data, implementing stochastic gradient descent (SGD) is helpful. Rewrite $\mathbf{X}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in terms of its n row vectors. The gradient becomes

$$\mathbf{u} = \mathbf{X}^T(\mathbf{y} - \boldsymbol{\pi}) = \sum_{i=1}^{n}(y_i - \pi_i)\mathbf{x}_i.$$

Thus the SGD incremental update can be as follows:

$$\boldsymbol{\beta}^{(k+1)} := \boldsymbol{\beta}^{(k)} + \alpha\,(y_i - \pi_i^{(k)})\mathbf{x}_i,$$

where α is the step size or learning rate.

The asymptotic variance-covariance matrix of the MLE $\widehat{\boldsymbol{\beta}}$ is given by

$$\operatorname{cov}\left(\widehat{\boldsymbol{\beta}}\right) = \left[\mathbf{X}^T \operatorname{diag}\{\pi_i(1-\pi_i)\}\mathbf{X}\right]^{-1}, \tag{29}$$

which can be estimated by substituting π_i with $\hat{\pi}_i$.

5.3 Example

To illustrate, we consider the kyphosis data (Chambers and Hastie, 1992) from a study of children who have had corrective spinal surgery. The data set contains 81 observations and four variables. A brief variable description is given below. The binary response, `kyphosis`, indicates whether kyphosis, a type of deformation was found on the child after the operation.

Table 2: Variable Description for the Kyphosis Data: Logistic Regression Example.

`kyphosis`	indicating if kyphosis is absent or present;
`age`	age of the child (in months);
`number`	number of vertebrae involved;
`start`	number of the first (topmost) vertebra operated on.

Logistic regression models can be fit using PROC LOGISTIC, PROC GLM, PROC CATMOD, and PROC GENMOD in SAS. In R, the function `glm` in the `base` library can be used. Another R implementation is also available in the package `Design`. In particular, function `lrm` provides penalized maximum likelihood estimation, i.e., the ridge estimator, for logistic regression.

Table 1 presents some selected fitting results from PROC LOGISTIC for model

$$\log\left\{\frac{\mathrm{P}(\texttt{kyphosis}=1)}{\mathrm{P}(\texttt{kyphosis}=0)}\right\} = \beta_0 + \beta_1 \cdot \texttt{age} + \beta_2 \cdot \texttt{number} + \beta_3 \cdot \texttt{start}.$$

Panel (a) gives the table of parameter estimates. The fitted logistic model is

$$\begin{aligned}\operatorname{logit}\{\mathrm{P}(\texttt{kyphosis}=1)\} &= -2.0369 + 0.0109 \times \texttt{age} + 0.4106 \times \texttt{number} \\ &\quad -0.2065 \times \texttt{start}.\end{aligned}$$

Accordingly, prediction of P(kyphosis = 1) can be obtained using (27). Panel (b) provides the estimates for the odds ratios (OR), $\exp(\hat{\beta}_j)$, and the associated 95% confidence intervals. The confidence interval for OR is constructed by taking the exponential of the lower and upper bounds of the confidence interval for β. Based on the results, we can conclude with 95% confidence that the odds of having kyphosis would be within $[0.712, 0.929]$ times if the number of the first (topmost) vertebra operated on, start, increases by one, for children with same fixed age and number values. Since 1 is not included in this confidence interval, the effect of start is significant at $\alpha = 0.05$, which is consistent with the Wald test in Panel (a). Panel (c) gives an example of the Wald test in its more general form for testing $H_0 : \beta_1 = \beta_2 = 0$. In this case,

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

of rank 2 when applying equation (20). The calculation also involves the estimated variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$, as shown in Panel (d). Referred to $\chi^2(2)$, the resultant p-value, 0.0804, is rather marginal. At the significance level $\alpha = 0.05$, one might consider dropping both age and number from the model.

6 Poisson Regression for Count Data

We next study the loglinear or Poisson models for count data as another example of GLM. Counts are frequencies of some event. Examples include the number of car accidents in different cities over a given period of time, the number of daily phone calls received in a call center, the number of students graduated from a high school, and so on. Count data are also commonly encountered in contingency tables.

6.1 The Loglinear Model

The Poisson or log-linear model is a popular GLM for count data, ideally when successive events occur independently and at the same rate. When the response Y_i follows Poisson(μ_i) for $i = 1, \ldots, n$, its probability distribution function is

$$\begin{aligned} f_{Y_i}(y_i) &= e^{-\mu_i} \mu_i^{y_i} / y_i! = \exp\left\{ y_i \log \mu_i - \mu_i - \log y_i! \right\} \\ &= \exp\left\{ \frac{y_i \theta_i - \exp(\theta_i)}{1} + (-\log y_i!) \right\} \end{aligned}$$

with the natural parameter $\theta_i = \log(\mu_i)$. Thus, in exponential family form, $b(\theta_i) = \exp(\theta_i)$, $a(\phi) = 1$, and $c(y_i, \phi) = -\log(y_i!)$. It follows from (3) that

$$\begin{aligned} E(Y_i) &= b'(\theta_i) = \exp(\theta_i) = \mu_i \\ \text{var}(Y_i) &= b''(\theta_i) = \exp(\theta_i) = \mu_i. \end{aligned}$$

The log-linear model is specified as

$$\log(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} \text{ for } i = 1, \ldots, n. \tag{30}$$

The canonical link, i.e., the logarithm function, is applied. In this model, one-unit increase in X_j has a multiplicative impact $\exp(\beta_j)$ on the mean response, holding other predictors fixed. It is worth mentioning that the log-linear or Poisson model in (30) is different from the ordinary Gaussian linear model with logarithm transformation on the response

$$\log y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \text{ with } \varepsilon_i \sim N(0, \sigma^2).$$

The log likelihood for the Poisson model (30) is given as

$$L(\boldsymbol{\beta}) = \sum_i y_i \log \mu_i - \sum_i \mu_i - \sum_i \log y_i!.$$

The last term $\sum_i \log y_i!$ can be ignored as it does not involve any parameter. It follows from (6), (7), and (10) that the gradient and the Fisher's information matrix are

$$\begin{aligned} \mathbf{u} &= \frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}) \\ \mathcal{I} &= \mathbf{X}^T \text{diag}(\boldsymbol{\mu}) \mathbf{X} = \left(\sum_{i=1}^n \mu_i x_{ij} x_{ij'} \right), \end{aligned}$$

where $\boldsymbol{\mu} = (\mu_i)$ denote the mean response vector and $\mathbf{W} = \text{diag}(\boldsymbol{\mu})$ is a diagonal matrix with diagonal elements μ_i. Therefore, in the iterative reweighted least squares algorithm for Fisher scoring, the intermediate response $\mathbf{z}^{(k)}$ in (17) has components

$$z_i^{(k)} = \log \mu_i^{(k)} + \frac{y_i - \mu_i^{(k)}}{\mu_i^{(k)}}.$$

The resulting MLE $\widehat{\boldsymbol{\beta}}$ has asymptotic variance-covariance matrix

$$\text{cov}\left(\widehat{\boldsymbol{\beta}}\right) = \left\{ \mathbf{X}^T \text{diag}(\boldsymbol{\mu}) \mathbf{X} \right\}^{-1}.$$

6.2 Example

We consider a school attendance data from Aitkin (1978), in which 146 children from Walgett, New South Wales, Australia, were classified by Culture, Age, Sex and Learner status. The response variable (Y) is the number of days absent from school in a particular school year. A brief description of the variables is provided in Table 4.

The objective is to explore the relationship between the four categorical predictors and school absence. To account for the four levels of `Age`, three dummy variables $\{Z_1^{\texttt{Age}}, Z_2^{\texttt{Age}}, Z_3^{\texttt{Age}}\}$ are introduced using the reference cell coding scheme such that

$$\begin{aligned} Z_1^{\texttt{Age}} &= \begin{cases} 1 & \text{if the child is in the "F3" age group;} \\ 0 & \text{otherwise.} \end{cases} \\ Z_2^{\texttt{Age}} &= \begin{cases} 1 & \text{if the child is in the "F2" age group;} \\ 0 & \text{otherwise.} \end{cases} \\ Z_3^{\texttt{Age}} &= \begin{cases} 1 & \text{if the child is in the "F1" age group;} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The other three predictors $\{\texttt{Sex}, \texttt{Lrn}, \texttt{Days}\}$ are all binary and 0-1 coded. The specific coding information for them is given in Panel (a) of Table 5. We consider the following log-linear model

$$\begin{aligned} \log(\texttt{days}) &= \beta_0 + \beta_1 \cdot \texttt{Eth}\beta_2 \cdot \texttt{Sex} + \beta_{31} \cdot Z_1^{\texttt{Age}} + \beta_{32} \cdot Z_2^{\texttt{Age}} \\ &\quad + \beta_{33} \cdot Z_3^{\texttt{Age}} + \beta_4 \cdot \texttt{Lrn}. \end{aligned}$$

Panel (b) of Table 5 presents the fitting results from PROC GENMOD. The ML estimates $\widehat{\beta}$s are shown in the first column, followed by their standard errors, the Wald 95% confidence intervals, and the χ^2 test of H_0: $\beta_j = 0$ for each of the individual parameters.

Therefore, given a set of predictor values, the predicted response can be obtained by equation

$$\begin{aligned} \widehat{\mu} &= \exp\Big\{2.7154 - 0.5336 \cdot \texttt{Eth} + 0.1616 \cdot \texttt{Sex} + 0.4277 \cdot Z_1^{\texttt{Age}} \\ &\quad + 0.2578 \cdot Z_2^{\texttt{Age}} - 0.3339 \cdot Z_3^{\texttt{Age}} + 0.3489 \cdot \texttt{Lrn}.\Big\} \end{aligned}$$

The model can be interpreted in the following way. Take the slope estimate for Sex for example. One may make the following statement: given boys and girls who are of the same ethnicity, same age, same learning status, the average number of absence days for boys is estimated to be $\exp(0.1616) = 1.1754$ times of that for girls, associated with a 95% confidence interval $\{\exp(0.0782), \exp(0.2450)\} = (1.0813, 1.2776)$.

References

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, **1**: 302–332.

Hastie, T., Tibshirani, R., Friedman, J. H. (2008). *Elements of Statistical Learning*, 2nd Edition. Springer, 2008.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logisitc Regression*. New York: John Wiley & Sons, Inc.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

Nelder, J. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of Royal Statistical Society, Series B*, **135**: 370–384.

Wang, H. and Leng, C. (2007). Unified LASSO Estimation by Least Squares Approximation. *Journal of the American Statistical Association*, **102**: 1039–1048.

Table 3: Analysis Results for the Kyphosis Data from PROC LOGISTIC.

(a) Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	−2.0369	1.4496	1.9744	0.1600
Age	1	0.0109	0.00645	2.8748	0.0900
Number	1	0.4106	0.2249	3.3340	0.0679
Start	1	−0.2065	0.0677	9.3045	0.0023

(b) Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.011	0.998	1.024
Number	1.508	0.970	2.343
Start	0.813	0.712	0.929

(d) Wald Test for $H_0 : \beta_1 = \beta_2 = 0$.

Wald Chi-Square	DF	Pr > ChiSq
5.0422	2	0.0804

(c) Estimated Covariance Matrix

	Intercept	Age	Number	Start
Intercept	2.101364	−0.00433	−0.27646	−0.0371
Age	−0.00433	0.000042	0.000337	−0.00012
Number	−0.27646	0.000337	0.050565	0.001681
Start	−0.0371	−0.00012	0.001681	0.004583

Table 4: Variable Description for the Log-Linear Regression Example.

Variable	**Description**	**Levels**
`Eth`	Ethnic background	Aboriginal ("A") or Not ("N")
`Sex`	Sex	"F" or "M"
`Age`	Age group	"F0", "F1", "F2", or "F3"
`Lrn`	Learner status: Average or Slow	"AL" or "SL"
`Days`	Days absent from school in the year	Integer

Table 5: Analysis Results for the School Absence Data from PROC GENMOD.

(a) Class Level Information

Class	Value	Design Variables		
`Eth`	N	1		
	A	0		
`Sex`	M	1		
	F	0		
`Age`	F3	1	0	0
	F2	0	1	0
	F1	0	0	1
	F0	0	0	0
`Lrn`	SL	1		
	AL	0		

(b) Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square
`Intercept`		1	2.7154	0.0647	2.5886	2.8422	1762.30
`Eth`	N	1	−0.5336	0.0419	−0.6157	−0.4515	162.32
`Sex`	M	1	0.1616	0.0425	0.0782	0.2450	14.43
`Age`	F3	1	0.4277	0.0677	0.2950	0.5604	39.93
	F2	1	0.2578	0.0624	0.1355	0.3802	17.06
	F1	1	−0.3339	0.0701	−0.4713	−0.1965	22.69
`Lrn`	SL	1	0.3489	0.0520	0.2469	0.4509	44.96