CHAPTER 4

GENERALIZED LINEAR MODELS

4.1 THE EXPONENTIAL-DISPERSION FAMILY

The pdf for the distributions in the **Exponential-Dispersion** family in their canonical form, with **naive** parameter θ_0, is of the form

$$f_{Y|\theta,\phi}(y;\theta_0,\phi) = \exp\left\{\frac{a_0(\theta_0)b_0(y)+c_0(\theta_0)}{r_0(\phi)} + d_0(y,\phi)\right\} \tag{4.1.1}$$

where $\phi > 0$ is a **dispersion** parameter and $c_0(.), d_0(.)$ are general scalar functions. In its **canonical parameterization** (by reparameterizing (4.1.1) writing $a_0(\theta_0) \to \theta$ so that θ is the **canonical parameter**), we express this

$$f_{Y|\theta,\phi}(y;\theta,\phi) = \exp\left\{\frac{\theta b(y)+c(\theta)}{r(\phi)} + d(y,\phi)\right\}. \tag{4.1.2}$$

The score function random variable is

$$U_\theta(Y) = \frac{\partial}{\partial\theta}\log f_{Y|\theta,\phi}(Y;\theta,\phi) = \frac{Y+c'(\theta)}{r(\phi)}$$

and taking expectation we have the usual results

$$E_{f_{Y|\theta,\phi}}[Y] = -c'(\theta) \qquad Var_{f_{Y|\theta,\phi}}[Y] = -c''(\theta)r(\phi)$$

Thus the variance is a function of the mean, and a function of the dispersion; in fact

$$VARIANCE = Function\ of\ \theta \times Function\ of\ \phi$$

and also the **variance is a function of the mean**; generically, we can write, for some function $V(.)$

$$E_{f_{Y|\theta,\phi}}[Y] = \mu \qquad Var_{f_{Y|\theta,\phi}}[Y] = V(\mu) \tag{4.1.3}$$

Example 4.1.1 Poisson Model

$$f_{Y|\lambda,\phi}(y;\lambda_0,\phi) = \frac{e^{-\lambda_0}\lambda_0^y}{y!} = \exp\left\{\frac{y\log\lambda_0-\lambda_0}{1} - \log y!\right\}$$

so that

$$a_0(\lambda_0) = \log\lambda_0 \qquad b_0(y) = y \qquad c_0(\lambda_0) = -\lambda_0 \qquad d_0(y,\phi) = -\log y! \qquad r_0(\phi) = 1$$

Thus we have canonical parameter $\lambda = \log\lambda_0$ and

$$f_{Y|\lambda,\phi}(y;\lambda,\phi) = \frac{\exp\{-\exp\{\lambda\}\}\exp\{\lambda\}^y}{y!} = \exp\left\{\frac{y\lambda-\exp\{\lambda\}}{r(\phi)} - \log y!\right\} \tag{4.1.4}$$

so that

$$b_0(y) = y \qquad c(\lambda) = -\exp\{\lambda\} \qquad d(y,\phi) = -\log y! \qquad r(\phi) = 1$$

Checking the expectation/variance results

$$E_{f_{Y|\lambda,\phi}}[Y] = -c'(\lambda) = -(-\exp\{\lambda\}) = \exp\{\lambda\} = \lambda_0$$

$$Var_{f_{Y|\theta,\phi}}[Y] = -c''(\lambda) r(\phi) = -(-\exp\{\lambda\}) \times 1 = \exp\{\lambda\} = \lambda_0$$

as expected; here we have by inspecting (4.1.3) variance function V defined by

$$V(t) = t.$$

Example 4.1.2 Binomial Model

$$f_{Y|\theta,\phi}(y;\theta_0,\phi) = \binom{n}{y}\theta_0^y (1-\theta_0)^{n-y} = \exp\left\{y \log\left(\frac{\theta_0}{1-\theta_0}\right) + n\log(1-\theta_0) + \log\binom{n}{y}\right\}$$

so that

$$a_0(\theta_0) = \log\left(\frac{\theta_0}{1-\theta_0}\right) \qquad b_0(y) = y \qquad c_0(\theta_0) = n\log(1-\theta_0) \qquad d_0(y,\phi) = \log\binom{n}{y} \qquad r(\phi) = 1$$

Thus we have canonical parameter

$$\theta = \log\left(\frac{\theta_0}{1-\theta_0}\right) \Longleftrightarrow \theta_0 = \frac{\exp\{\theta\}}{1+\exp\{\theta\}}$$

and

$$\begin{aligned} f_{Y|\theta,\phi}(y;\theta,\phi) &= \binom{n}{y}\left(\frac{\exp\{\theta\}}{1+\exp\{\theta\}}\right)^y \left(\frac{1}{1+\exp\{\theta\}}\right)^{n-y} \\ &= \exp\left\{\frac{y\theta - n\log(1+\exp\{\theta\})}{r(\phi)} + \log\binom{n}{y}\right\} \end{aligned} \tag{4.1.5}$$

so that

$$b(y) = y \qquad c(\theta) = -n\log(1+\exp\{\theta\}) \qquad d(y,\phi) = \log\binom{n}{y}.$$

Checking the expectation/variance results

$$E_{f_{Y|\theta,\phi}}[Y] = -c'(\theta) = -\left(-\frac{n\exp\{\theta\}}{(1+\exp\{\theta\})}\right) = n\theta_0$$

$$Var_{f_{Y|\theta,\phi}}[Y] = -c''(\theta) r(\phi) = -\left(-\frac{n\exp\{\theta\}}{(1+\exp\{\theta\})^2}\right) = n\theta_0(1-\theta_0)$$

as expected; here we have by inspecting (4.1.3) variance function V defined as follows $\mu = n\theta_0$

$$\Longrightarrow \quad Var_{f_{Y|\theta,\phi}}[Y] = n\theta_0(1-\theta_0) = \frac{\mu(n-\mu)}{n}$$

$$\Longrightarrow \quad V(t) = \frac{t(n-t)}{n}.$$

Example 4.1.3 Binomial Model: reformulation
It is common in the binomial case to regard the response as $m = y/n$ instead of y. Thus we have, in place of (4.1.5),

$$\begin{aligned} f_{M|\theta,\phi}(m;\theta,\phi) &= \binom{n}{nm}\left(\frac{\exp\{\theta\}}{1+\exp\{\theta\}}\right)^{nm}\left(\frac{1}{1+\exp\{\theta\}}\right)^{n-nm} \\ &= \exp\left\{\frac{m\theta-\log\left(1+\exp\{\theta\}\right)}{r\left(\phi\right)}+\log\binom{n}{nm}\right\} \end{aligned} \tag{4.1.6}$$

where now

$$c(\theta) = -\log\left(1+\exp\{\theta\}\right) \qquad r\left(\phi\right) = \frac{1}{n}$$

In this reformulation

$$\begin{aligned} E_{f_{M|\theta,\phi}}[M] &= -c'(\theta) = -\left(-\frac{\exp\{\theta\}}{\left(1+\exp\{\theta\}\right)}\right) = \theta_0 \\ Var_{f_{M|\theta,\phi}}[M] &= -c''(\theta)r\left(\phi\right) = -\left(-\frac{\exp\{\theta\}}{\left(1+\exp\{\theta\}\right)^2}\right)\frac{1}{n} = \frac{\theta_0\left(1-\theta_0\right)}{n} \end{aligned}$$

so that, now $\mu = \theta_0$

$$\Longrightarrow Var_{f_{M|\theta,\phi}}[M] = \frac{\theta_0\left(1-\theta_0\right)}{n} = \mu\left(1-\mu\right)r\left(\phi\right)$$

This simplifies some of the calculations below considerably, as we now have

$$Var_{f_{M|\theta,\phi}}[M] = Function\ of\ \theta \times Function\ of\ \phi$$

as required above for independent modelling of mean and dispersion.

4.2 THE GENERALIZED LINEAR MODEL

The central idea of **Generalized Linear Models** (GLMs) is to extend the ideas from the normal model to allow the possibility of modelling non-normal data. In the GLM, we will model

$$E_{f_{Y|X}}[Y_i|X_i = x_i] = g^{-1}\left(x_i^T\beta\right) \qquad \text{where} \qquad x_i^T\beta = \beta_0 + \sum_{j=1}^{K}\beta_j x_{ij}, \text{ say}$$

for some monotonic/invertible function g; in the normal linear model, g is the **identity** function.

4.2.1 GLM TERMINOLOGY

There are two key terms in the model description:

- **Linear predictor**: for observed predictor $x_i = (x_{i1}, ..., x_{iK})$ and parameters $\beta = (\beta_0, \beta_1, ...\beta_K)$, the **linear predictor** is

$$\eta_i = x_i^T\beta = \beta_0 + \sum_{j=1}^{K}\beta_j x_{ij}$$

- **Link function:** a **link function** g is a function that connects the linear predictor to the expected value of the response

$$g\left(E_{f_{Y|X}}[Y_i|X_i = x_i]\right) = x_i^T\beta.$$

4.2.2 CANONICAL AND OTHER LINK FUNCTIONS

For the **Poisson** model;

- The **canonical link** is the **log** link: it is the link function that connects the expectation (in this case, the naive parameter λ_0) to the linear predictor

  $$\lambda = \log \lambda_0 = x^T \beta$$

  Here,

  $$E_{f_{Y|\lambda}}[Y] = \lambda_0 = \exp\{\lambda\} \qquad \therefore \qquad g\left(E_{f_{Y|\lambda}}[Y]\right) = x^T\beta \qquad \text{where } g(t) = \log t$$

- **Power link**

  $$g(t) = t^\alpha$$

  for some real parameter α

- **Box-Cox link**

  $$g(t) = \frac{t^\alpha - 1}{\alpha}$$

  for some real parameter α.

For the **Binomial** model;

- The **canonical link** is the **logit** link connects the expectation (again, naive parameter θ_0):

  $$\theta = \log\left(\frac{\theta_0}{1-\theta_0}\right) = x^T\beta$$

  Here,

  $$E_{f_{M|\lambda}}[M] = \theta_0 = \frac{\exp\{\theta\}}{(1+\exp\{\theta\})} \qquad \therefore \qquad g\left(E_{f_{m|\lambda}}[M]\right) = x^T\beta \qquad \text{where } g(t) = \log\left(\frac{t}{1-t}\right)$$

- **Probit link**

  $$g(t) = \Phi^{-1}\left(\frac{\exp\{t\}}{1+\exp\{t\}}\right) = x^T\beta \qquad \therefore \qquad \Phi^{-1}(\theta_0) = x^T\beta$$

  where Φ is the standard normal cdf.

- **Complementary log-log link**

  $$g(t) = \log\{\log(1+\exp\{t\})\} = x^T\beta \qquad \therefore \qquad \log\{-\log(1-\theta_0)\} = x^T\beta$$

- **Log-log link**

  $$g(t) = -\log\{-t + \log(1+\exp\{t\})\} = x^T\beta \qquad \therefore \qquad -\log\{-\log\theta_0\} = x^T\beta$$

4.3 CHECKING THE FIT OF A GLM

In the normal linear framework, the fit of a model is assessed by inspection of the magnitude of the residual sum of squares (RSS) and comparing it with the fitted sum of squares (FSS) in an ANOVA F-test. For example, to test a model with K predictors plus an intercept ($K+1$ parameters) against the model with just an intercept (1 parameter), we use either a Chi-squared statistic

$$\frac{\left[S\left(\widehat{\beta}_0\right) - S\left(\widehat{\beta}\right)\right]}{\sigma^2} \sim \chi^2_K$$

or the F-statistic,

$$\frac{\left[S\left(\widehat{\beta}_0\right) - S\left(\widehat{\beta}\right)\right]/K}{S\left(\widehat{\beta}\right)/(n-K-1)} \sim Fisher\left(K, n-K-1\right)$$

or by inspecting the error in fit as measured by the **residual,** $e = y_i - \widehat{y}_i = y_i - x_i^T\widehat{\beta}$, where $\widehat{\beta}_0$ and $\widehat{\beta}$ are mles computed under the 1 and $K+1$ parameter models respectively. In the GLM case, we will use similar, Likelihood Ratio (LR) statistics to perform tests.

4.3.1 DEVIANCE

In the following we use the following notation for data Y modelled via linear predictor $\eta = x^T\beta$ through canonical parameter θ and related expected value $\mu = E_{f_{Y|X,\beta}}[Y]$ with link function, g. After the model is fitted, we have ML estimates in the linear predictor, β, for the following parameters

$$\widehat{\eta} = x^T\widehat{\beta} \qquad \widehat{\theta} = g^{-1}\left(x^T\widehat{\beta}\right) \qquad \widehat{\mu} = h\left(\widehat{\theta}\right)$$

The value $\hat{\mu}$ is the expected value of Y under the fitted model, and the function h is the function that maps θ to expectation μ. We may also write

$$\widehat{y} = \widehat{\mu}$$

The goodness of fit of a GLM is measured in terms of **deviance**. From a previous definition, the deviance, D, for a model M is the likelihood ratio statistic in an LR test of the model against the **saturated** model, S,

$$D = 2\log\frac{l_S\left(\widehat{\beta}_S\right)}{l_M\left(\widehat{\beta}_M\right)} = -2\log\frac{l_M\left(\widehat{\beta}_M\right)}{l_S\left(\widehat{\beta}_S\right)}$$

- $\widehat{\beta}_M$ is the mle under model M
- $\widehat{\beta}_S$ is the mle baseline model the saturated model, which corresponds to the **best possible fit,** which has the same number of parameters as data points, and which occurs when

$$\widehat{y}_i = \widehat{\mu}_i = y_i$$

- l_M and l_S are the likelihood functions under the model and saturated model respectively.

Example 4.3.1 BERNOULLI MODEL

- **MODEL:** Single predictor, X, canonical logit link

$$\eta_i = \log\left(\frac{\theta_i}{1-\theta_i}\right) = x_i^T\beta = \beta_0 + \beta_1 x_i$$

so that

$$\widehat{\theta}_i = \frac{\exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right\}}{1+\exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right\}}.$$

Thus

$$\begin{aligned} l_M\left(\widehat{\beta}_M\right) &= \prod_{i=1}^{n} \widehat{\theta}_i^{y_i}\left(1-\widehat{\theta}_i\right)^{1-y_i} \qquad (4.3.7)\\ &= \prod_{i=1}^{n}\left(\frac{\exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right\}}{1+\exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right\}}\right)^{y_i}\left(\frac{1}{1+\exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right\}}\right)^{1-y_i} \end{aligned}$$

- **SATURATED MODEL:** Must have

$$\widehat{y}_i = \widehat{\theta}_i = y_i \qquad (4.3.8)$$

- **DEVIANCE:** From (4.3.7) and (4.3.8)

$$D = -2\log\frac{l_M\left(\widehat{\beta}_M\right)}{l_S\left(\widehat{\beta}_S\right)} = -2\sum_{i=1}^{n} y_i \log\left(\frac{\widehat{\theta}_i}{y_i}\right) + (1-y_i)\log\left(\frac{1-\widehat{\theta}_i}{1-y_i}\right)$$

But here

$$l_S\left(\widehat{\beta}_S\right) = \prod_{i=1}^{n} y_i^{y_i}\left(1-y_i\right)^{1-y_i} = 1$$

so that

$$D = -2\sum_{i=1}^{n} y_i \log\left(\widehat{\theta}_i\right) + (1-y_i)\log\left(1-\widehat{\theta}_i\right)$$

Example 4.3.2 POISSON MODEL

- **MODEL:** Single predictor, X, canonical log link

$$\eta_i = \log\lambda_i = x_i^T\beta = \beta_0 + \beta_1 x_i$$

so that

$$\widehat{\lambda}_i = \exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right\}.$$

Thus

$$l_M\left(\widehat{\beta}_M\right) = \prod_{i=1}^{n} \frac{\exp\left\{-\widehat{\lambda}_i\right\}\widehat{\lambda}_i^{y_i}}{y_i!} \tag{4.3.9}$$
$$= \prod_{i=1}^{n_i} \exp\left\{-\exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right\}\right\} \exp\left\{y_i\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right\} \times \frac{1}{\left\{\prod_{i=1}^{n} y_i!\right\}}$$

and hence

$$\log l_M\left(\widehat{\beta}_M\right) = \sum_{i=1}^{n}\left[y_i\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) - \exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right\} - \log y_i!\right]$$

- **SATURATED MODEL:** Must have

$$\widehat{y}_i = \widehat{\lambda}_i = y_i \tag{4.3.10}$$

- **DEVIANCE:** From (4.3.9) and (4.3.10)

$$D = -2\log\frac{l_M\left(\widehat{\beta}_M\right)}{l_S\left(\widehat{\beta}_S\right)} = -2\sum_{i=1}^{n}\left[y_i\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i - \log y_i\right) - \left(\exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right\} - y_i\right)\right] \tag{4.3.11}$$

4.3.2 THE NULL MODEL AND DEVIANCE

The **null model** is the model where $\mu_i = \mu$ (a constant) for all $i = 1, ..., n$. For example, in the binomial model with canonical logistic link, the null model specifies $\theta_i = \theta$, and

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 \Leftrightarrow \theta_i = \frac{\exp\{\beta_0\}}{1+\exp\{\beta_0\}}$$

whereas in the Poisson model, the null model $\lambda_i = \lambda$, and

$$\log\lambda_i = \beta_0 \Leftrightarrow \lambda_i = \exp\{\beta_0\}$$

where, in both cases β_0 is a scalar parameter.

Thus we have a complete range of model fits to calibrate the fit of any individual model:

$$\text{SATURATED MODEL} \rightarrow \text{MODEL} \rightarrow \text{NULL MODEL}$$

$$\text{MOST COMPLEX} \rightarrow \text{LEAST COMPLEX}$$

$$\text{LOWEST DEVIANCE} \rightarrow \text{HIGHEST DEVIANCE}$$

4.3.3 DEVIANCE AND SCALED DEVIANCE

In the Exponential-Dispersion family, where for data point i

$$f_{Y_i|\theta,\phi}(y_i;\theta_i,\phi) = \exp\left\{\frac{\theta_i b\left(y_i\right)+c(\theta_i)}{r_i(\phi)}+d\left(y_i,\phi\right)\right\}$$

let

- MODEL fit : $\widehat{\theta}_i$ defined by

$$\widehat{\mu}_i = -c\left(\widehat{\theta}_i\right) = g^{-1}\left(\widehat{\eta}_i\right) = g^{-1}\left(x_i^T\widehat{\beta}_M\right)$$

- SATURATED fit : $\widetilde{\theta}_i$ defined by

$$\widetilde{\mu}_i = -c\left(\widetilde{\theta}_i\right) = y_i$$

- w_i is defined by

$$w_i = \frac{\phi}{r_i\left(\phi\right)}.$$

Then the **scaled deviance**, D^*, is defined by

$$D^* = \sum_{i=1}^{n}\frac{2w_i}{\phi}\left\{b(y_i)\left(\widetilde{\theta}_i-\widehat{\theta}_i\right)+\left(c\left(\widetilde{\theta}_i\right)-c\left(\widehat{\theta}_i\right)\right)\right\} = \frac{D}{\phi}$$

and we have in full

$$D^*\left(y;\widehat{\theta},\phi\right) = \frac{D\left(y;\widehat{\theta}\right)}{\phi}$$

4.3.4 STATISTICAL PROPERTIES OF DEVIANCE

Likelihood Ratio theory gives a means of calibrating the magnitude of the deviance; we have, by the usual asymptotic theory of MLEs,

$$D^*\left(y;\widehat{\theta},\phi\right) = \frac{D\left(y;\widehat{\theta}\right)}{\phi} \overset{A}{\sim} \chi^2_{n-K-1} \tag{4.3.12}$$

if η has $K+1$ parameters. From this result we have two possible estimates of dispersion parameter ϕ; the **Deviance-based** estimate

$$\widehat{\phi}_D = \frac{D\left(y;\widehat{\theta}\right)}{n-K-1}$$

or the **Pearson-type** estimate

$$\widehat{\phi}_P = \frac{1}{n-K-1}\sum_{i=1}^{n}\frac{\left(y_i-\widehat{\mu}_i\right)^2}{V\left(\widehat{\mu}_i\right)/w_i}$$

For choosing between models M_1 (K_1 predictors, parameter θ_1) and M_2 (K_2 predictors, parameter θ_2), where $K_2 > K_1$, we have

$$\frac{D_{M_1}\left(y;\widehat{\theta}_1\right) - D_{M_2}\left(y;\widehat{\theta}_2\right)}{\phi} \overset{A}{\sim} \chi^2_{K_2-K_1}$$

and, if ϕ is not known

$$\frac{\left(D_{M_1}\left(y;\widehat{\theta}_1\right) - D_{M_2}\left(y;\widehat{\theta}_2\right)\right)/(K_2-K_1)}{\widehat{\phi}} \overset{A}{\sim} Fisher\left(K_2-K_1, n-K_2-1\right)$$

where $\widehat{\phi}$ is either of $\widehat{\phi}_D$ or $\widehat{\phi}_P$.

NOTE : These results are only approximate, and can often be poor, especially if the number of parameters is large relative to the number of data points. The distributional result in (4.3.12) is conditional on the **TRUE** value of θ, indicates that the estimator given by $D^*\left(y;\widehat{\theta},\phi\right)$ has the same approximate chi-squared distribution whatever the value of θ. This is often demonstrably not true; the MLEs themselves are often biased in such "non-regular" problems.

4.3.5 RESIDUALS

Three types of residual are used for checking the fit of a GLM

- **Pearson residual**

$$e_i = \frac{y_i - \widehat{y}_i}{\sqrt{V\left(\widehat{y}_i\right)}}$$

- **Anscombe residual**

$$e_{Ai} = \frac{A\left(y_i\right) - A\left(\widehat{y}_i\right)}{A'\left(\widehat{y}_i\right)\sqrt{V\left(\widehat{y}_i\right)}}$$

  where function A is defined by

$$A(t) = \int_0^t \frac{1}{V^{1/3}(s)} ds$$

- **Deviance Residual**

$$D\left(y;\widehat{\theta}\right) = \sum_{i=1}^{n} D_i\left(y_i;\widehat{\theta}\right) = \sum_{i=1}^{n} D_i$$

  where

$$D_i = 2w_i\left\{b\left(y_i\right)\left(\widetilde{\theta} - \widehat{\theta}_i\right) + \left(c\left(\widetilde{\theta}_i\right) - c\left(\widehat{\theta}_i\right)\right)\right\}$$

  Then the deviance residual is

$$e_{Di} = sign\left(y_i - \widehat{y}_i\right)\sqrt{D_i} = \begin{cases} \sqrt{D_i} & y_i > \widehat{y}_i \\ -\sqrt{D_i} & y_i < \widehat{y}_i \end{cases}$$

  In the Poisson case, from (4.3.11)

$$e_{Di} = sign\left(y_i - \widehat{y}_i\right)\left\{2\left(y_i \log\frac{y_i}{\widehat{y}_i} - y_i + \widehat{y}_i\right)\right\}^{1/2}$$

  where

$$\widehat{y}_i = \widehat{\lambda}_i = \exp\left\{x_i^T\beta\right\}$$

4.4 COUNTING THE NUMBERS OF PARAMETERS IN A MODEL

Each term in the linear or generalized model formulation contributes to the **model complexity**, that is, the number of parameters fitted by the model. The model complexity is a key component in assessing the worth of a model

4.4.1 CONTINUOUS PREDICTORS

Continuous predictors contribute one parameter (or coefficient) to the fit of the model when the appear as main effects, and one parameter for each separate interaction term fitted. Recall the extended normal linear model linear predictor form:

$$x_i^T \beta = \beta_0 + \sum_{j=1}^{K} \beta_j g(x_{ij}).$$

The number of parameters fitted is $(K + 1)$. The interaction case is described below.

4.4.2 DISCRETE PREDICTORS

For two discrete predictor (**factors**), fitting the

- A having I levels
- B having J levels

Models that can be fitted are:

1. A

$$\begin{aligned} \eta &= \mu && i = 1 \\ &= \mu + \alpha_i && i = 2, ..., I \end{aligned}$$

Number of parameters is

$$1 + (I - 1) = I$$

that is, μ, plus $(I - 1)$ α parameters.

2. $A + B$

$$\begin{aligned} \eta &= \mu && i = 1, j = 1 \\ &= \mu + \alpha_i && i = 2, ..., I, j = 1 \\ &= \mu + \beta_j && i = 1, j = 2, ..., J \\ &= \mu + \alpha_i + \beta_j && i = 2, ..., I, j = 2, ..., J \end{aligned}$$

Number of parameters is

$$1 + (I - 1) + (J - 1) = I + J - 1$$

that is, μ, plus $(I - 1)$ α parameters, plus $(J - 1)$ β parameters.

3. $A * B$

$$\begin{aligned}\eta &= \mu && i = 1, j = 1\\ &= \mu + \alpha_i && i = 2, ..., I, j = 1\\ &= \mu + \beta_j && i = 1, j = 2, ..., J\\ &= \mu + \alpha_i + \beta_j + \gamma_{ij} && i = 2, ..., I, j = 2, ..., J\end{aligned}$$

Number of parameters is

$$1 + (I-1) + (J-1) + (I-1)(J-1) = IJ$$

that is, μ, plus $(I-1)$ α parameters, plus $(J-1)$ β parameters plus $(I-1)(J-1)$ γ parameters.

- μ is the **GRAND MEAN**
- α_i are the **ROW EFFECTS**
- β_j are the **COLUMN EFFECTS**

4.4.3 MIXED TERMS

It is possible to combine continuous and discrete predictors; suppose for example we have a continuous predictor X and a discrete predictor A having I levels (labelled $i = 1, 2, ...$). Then the models that can be fitted are as follows

- X

$$\eta = x^T\beta = \beta_0 + \beta_1 x$$

which is a model with 2 parameters.

- A

$$\begin{aligned}\eta &= \mu && i = 1\\ &= \mu + \alpha_i && i = 2, ..., I\end{aligned}$$

which is a model with $1 + (I-1) = I$ parameters.

- $X + A$

$$\begin{aligned}\eta &= \mu + \beta_1 x && i = 1\\ &= \mu + \alpha_i + \beta_1 x && i = 2, ..., I\end{aligned}$$

which is a model with $2 + (I-1) = I + 1$ parameters, and where, $\mu = \beta_0$ and an alternative reparameterization is

$$\eta = \begin{cases} \beta_{10} + \beta_1 x & i = 1\\ \beta_{20} + \beta_1 x & i = 2\\ \vdots & \\ \beta_{I0} + \beta_1 x & i = I\end{cases}$$

where $\beta_{i0} = \mu + \alpha_i$. This model corresponds to having I parallel lines for the linear predictor

- $X * A$

$$\begin{aligned} \eta &= \mu + \beta_1 x && i = 1 \\ &= (\mu + \alpha_i) + (\beta_1 + \beta_{i1})\, x && i = 2, ..., I \end{aligned}$$

which is a model with $2 + 2 * (I - 1) = 2I$ parameters, and whereas alternative reparameterization is

$$\eta = \begin{cases} \beta_{10} + \beta_{11} x & i = 1 \\ \beta_{20} + \beta_{21} x & i = 2 \\ \vdots & \\ \beta_{I0} + \beta_{I1} x & i = I \end{cases}$$

4.5 EXAMPLES

4.5.1 ICU DATA

DATA: $n = 200$ Patients in Intensive Care Unit

RESPONSE: Survival for Duration of Follow-Up Period $Y_i = \begin{cases} 1 & \text{Died} \\ 0 & \text{Survived} \end{cases}$

PREDICTORS: Include

- **AGE :** continuous predictor (analysis omitted here)
- **SEX :**discrete factor, 2 levels (0=MALE,1=FEMALE)
- **RACE :** discrete factor, 3 levels (1=WHITE, 2=BLACK, 3=OTHER)

IN CATEGORY

		RACE		
		1	2	3
SEX	0	108	10	6
	1	67	5	4

DEATHS

		RACE		
		1	2	3
SEX	0	23	0	1
	1	14	1	1

that is, of the 108 patients who were MALE and WHITE, 23 died during follow up, and $108 - 23 = 75$ survived, and so on. Denote the elements in the left-hand table $\{n_{ij}, i = 0, 1 \text{ and } j = 1, 2, 3\}$ and in the right-hand table $\{s_{ij}, i = 0, 1 \text{ and } j = 1, 2, 3\}$. Hence, in the previous notation

$$I = 2 \qquad \text{and} \qquad J = 3$$

LIKELIHOOD: Full likelihood is

$$\prod_{i=0}^{1} \prod_{j=1}^{3} \left\{ \prod_{k \in C_{ij}} \theta_{ij}^{y_k} (1 - \theta_{ij})^{1 - y_k} \right\}$$

where C_{ij} is the set of patients in the cohort with SEX$= i$ and RACE$= j$, which simplifies to

$$\prod_{i=0}^{1} \prod_{j=1}^{3} \left\{ \theta_{ij}^{s_{ij}} (1 - \theta_{ij})^{n_{ij} - s_{ij}} \right\}$$

if only the to discrete predictors are fitted, where

$$\theta_{ij} = \mathrm{P}\left[\text{Death during follow-up for SEX} = i \text{ and RACE} = j\right]$$

LINK FUNCTION: logit

$$\text{logit}\left(\theta_{ij}\right) = \log\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right)$$

MODELS:

MODEL	TERMS	NUMBER OF PARAMETERS
NULL	1	1
MAIN EFFECTS	SEX	$1 + (I-1) = 1 + (2-1) = 2$
	RACE	$1 + (J-1) = 1 + (3-1) = 3$
	SEX+RACE	$1 + (I-1) + (J-1) = 1 + (2-1) + (3-1) = 4$
INTERACTION	SEX*RACE	$1 + (I-1) + (J-1) + (I-1) * (J-1)$ $= 1 + (2-1) + (3-1) + (2-1)(3-1) = 1 + 1 + 2 + 2 = 6$

ANALYSIS OF DEVIANCE:
Let

- p be the number of parameters in the model fitted
- DF be the (residual) degrees of freedom (note $p + DF = n$)
- D be the (residual, unscaled) deviance
- $\triangle DF$ be the difference in degrees of freedom compared with null model

$$\triangle DF = DF_{NULL} - DF_{MODEL}$$

- $\triangle D$ be the difference in deviance compared with null model

$$\triangle D = D_{NULL} - D_{MODEL}$$

- $\chi^2_{\triangle DF}\left(1-\alpha\right)$ is the $100(1-\alpha)\%$ quantile of the Chi-squared distribution with $\triangle DF$ degrees of freedom

MODEL	p	DF	D	$\triangle DF$	$\triangle D$	$\chi^2_{\triangle DF}\left(0.95\right)$
NULL	1	199	200.101	-	-	
SEX	2	198	200.076	1	0.025	3.8415
RACE	3	197	197.901	2	2.200	5.9915
SEX+RACE	4	196	197.836	3	2.265	7.8147
SEX*RACE	6	194	195.456	5	4.645	11.0705

CONCLUSION: No model fits significantly better than the null; neither SEX nor RACE is an important predictor for death during follow up.

FITTED VALUES:

Resulting model is

$$\text{logit}\,(\theta_{ij}) = \log\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) = \mu$$

and, using the following SPLUS commands

```
>fit.0<-glm(y~1,family=binomial,data=ICU.SUB)
>summary(fit.0)
Coefficients:
               Value   Std.Error  t value
(Intercept) -1.386294 0.1767767 -7.842065
```

$$\widehat{\mu} = -1.386294 \Longrightarrow \text{For each } (i,j)\,,\ \widehat{\theta}_{ij} = \text{expit}\,\{\widehat{\mu}\} = \frac{\exp\{\widehat{\mu}\}}{1+\exp\{\widehat{\mu}\}} = 0.2$$

Note that this is the usual MLE estimate resulting from these data assuming that there are no differences between subgroups defined by SEX and RACE cross-classifications, that is

$$\widehat{\theta}_{ij} = \widehat{\theta} = \frac{\sum_{i=0}^{1}\sum_{j=1}^{3} s_{ij}}{\sum_{i=0}^{1}\sum_{j=1}^{3} n_{ij}} = \frac{40}{200} = 0.2$$

4.5.2 LOW BIRTH WEIGHT DATA

DATA: $n = 189$ infants

RESPONSE: Birth-weight indicator LOW $Y_i = \begin{cases} 1 & \text{Birth weight } < 2500g \\ 0 & \text{Birth-weight } \geq 2500g \end{cases}$

PREDICTORS:

- **SMOKE :**discrete factor, 2 levels (0=NOT SMOKER,1=SMOKER)
- **RACE :** discrete factor, 3 levels (1=WHITE, 2=BLACK, 3=OTHER)

Raw counts

IN CATEGORY

		RACE		
		1	2	3
SMOKE	0	44	16	55
	1	52	10	12

LOW BIRTHWEIGHT

		RACE		
		1	2	3
SMOKE	0	4	5	20
	1	19	6	5

that is, of the 44 patients who were NOT SMOKERs and WHITE, 4 had low infant birthweight, and $44 - 4 = 40$ did not have low infant birthweight, and so on. Again,

$$I = 2 \qquad \text{and} \qquad J = 3.$$

MODELS:

MODEL	TERMS	NUMBER OF PARAMETERS
NULL	1	1
MAIN EFFECTS	SMOKE	$1+(I-1)=1+(2-1)=2$
	RACE	$1+(J-1)=1+(3-1)=3$
	SMOKE+RACE	$1+(I-1)+(J-1)=1+(2-1)+(3-1)=4$
INTERACTION	SMOKE*RACE	$1+(I-1)+(J-1)+(I-1)*(J-1)$ $=1+(2-1)+(3-1)+(2-1)(3-1)=1+1+2+2=6$

ANALYSIS OF DEVIANCE:

MODEL	p	DF	D	$\triangle DF$	$\triangle D$	$\chi^2_{\triangle DF}(0.95)$
NULL	1	188	234.6720	-	-	
SMOKE	2	187	229.8046	1	4.8674	3.8415
RACE	3	186	229.6616	2	5.0104	5.9915
SMOKE+RACE	4	185	219.9747	3	14.6973	7.8147
SMOKE*RACE	6	183	216.8178	5	17.8542	11.0705

CONCLUSION: SMOKE **and** RACE are **both** important predictors; although RACE is not important when fitted as a main effect on its own, the models

- SMOKE+RACE and
- SMOKE*RACE=SMOKE+RACE+SMOKE.RACE

both fit significantly better than the null model. We can deduce the best fitting model as follows: SMOKE is an important predictor, as

$$\triangle D_{SMOKE} = D_{NULL} - D_{SMOKE} = 4.8674 > 3.8415 = \chi^2_1(0.95)$$

and thus this is significant at the 0.05 level. In addition, comparing the model fits of the SMOKE and SMOKE+RACE model, we have

$$D_{SMOKE} - D_{SMOKE+RACE} = 229.8046 - 219.9747 = 9.8299$$

$$DF_{SMOKE} - DF_{SMOKE+RACE} = 187 - 185 = 2$$

so

$$D_{SMOKE} - D_{SMOKE+RACE} = 9.8299 > 5.9915 = \chi^2_2(0.95)$$

and thus SMOKE+RACE fits significantly better than the SMOKE model. Finally, comparing SMOKE+RACE and SMOKE*RACE model, we have

$$D_{SMOKE+RACE} - D_{SMOKE*RACE} = 219.9747 - 216.8178 = 3.1569$$

$$DF_{SMOKE+RACE} - DF_{SMOKE*RACE} = 185 - 183 = 2$$

so

$$D_{SMOKE+RACE} - D_{SMOKE*RACE} = 3.1569 \leq 5.9915 = \chi^2_2(0.95)$$

and thus SMOKE*RACE does not fit significantly better than the SMOKE+RACE model. Hence the preferred model is

SMOKE+RACE

Hence, a model summary for the linear predictors is as follows

		RACE		
		1	2	3
SMOKE	0	μ	$\mu+\beta_2$	$\mu+\beta_3$
	1	$\mu+\alpha_2$	$\mu+\alpha_2+\beta_2$	$\mu+\alpha_2+\beta_3$

PARAMETER ESTIMATES:

Resulting model is

$$\text{logit}\left(\theta_{ij}\right) = \log\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) = \eta_{ij}$$

and, using the following SPLUS commands

```
>fit.3<-glm(LOW~factor(SMOKE)+factor(RACE),family=binomial,data=LBW)
>summary(fit.3)
                 Value  Std.Error  t-value
 (Intercept) -0.5516383 0.1830937 -3.012873
factor(SMOKE) 0.5579517 0.1839633  3.032950
factor(RACE)1 0.5419994 0.2444412  2.217300
factor(RACE)2 0.1888159 0.1251242  1.509028
```

The default SPLUS parameterization is not the one implied above; to obtain the correct parameterization we must issue the commands

```
>options(contrasts=c(''contr.treatment'',''contr.poly''))
>fit.3<-glm(LOW~factor(SMOKE)+factor(RACE),family=binomial,data=LBW)
>summary(fit.3)
                 Value  Std.Error  t-value
 (Intercept) -1.840405 0.3508201 -5.246009
factor(SMOKE) 1.115903 0.3679267  3.032950
factor(RACE)2 1.083999 0.4888823  2.217300
factor(RACE)3 1.108447 0.3988288  2.779256
```

from which we obtain the parameter estimates

$$\widehat{\mu} = -1.840405 \qquad \widehat{\alpha}_2 = 1.115903 \qquad \widehat{\beta}_2 = 1.083999 \qquad \widehat{\beta}_3 = 1.108447$$

Carrying out the Wald-type test of the hypotheses for

$$\begin{aligned} H_0 &: \beta = 0 \\ H_1 &: \beta \neq 0 \end{aligned}$$

for general parameter β, based on the approximate result that, under H_0

$$\frac{\widehat{\beta}}{s.e.\left(\widehat{\beta}\right)} \overset{A}{\sim} N(0,1)$$

we can deduce from the `t-value` parameters in the table, we deduce that the parameters are all significantly different from zero.

4.6 ITERATIVELY REWEIGHTED LEAST SQUARES

In the canonical exponential-dispersion family with function b the identity, the log-likelihood takes the form

$$l(\theta,\phi)=\sum_{i=1}^{n}\frac{w_i\left(y_i\theta_i+c(\theta_i)\right)}{\phi}+\sum_{i=1}^{n}d\left(y_i,\frac{\phi}{w_i}\right) \tag{4.6.13}$$

for canonical parameter θ_i and, the expectation and variance of observable Y_i are

$$E[Y_i]=-\frac{\partial c(\theta)}{\partial\theta_i}=-c'(\theta_i)=\mu_i \qquad Var[Y_i]=-\frac{\phi}{w_i}\frac{\partial^2 c(\theta)}{\partial\theta_i^2}=-\frac{\phi}{w_i}c''(\theta_i)=\frac{\phi}{w_i}V(\mu_i) \tag{4.6.14}$$

The GLM with link function g sets

$$g(\mu_i)=g\left(-c'(\theta_i)\right)=x_i^T\beta=\eta_i \tag{4.6.15}$$

for a model with parameters $\beta=(\beta_1,...,\beta_K)$. With the canonical link g we have merely that

$$g\left(-c'(\theta_i)\right)=\theta_i$$

so that $\theta_i=x_i^T\beta$.

A set of K **score equations** is obtained by differentiation (4.6.13) partially with respect to the elements of β and equating to zero. For example, for β_k

$$\frac{\partial}{\partial\beta_k}\{l(\beta,\phi)\}=\sum_{i=1}^{n}\frac{w_i}{\phi}\left(y_i\frac{\partial\theta_i}{\partial\beta_k}+\frac{\partial c(\theta_i)}{\partial\beta_k}\right)=\sum_{i=1}^{n}\frac{w_i}{\phi}\left(y_i+\frac{\partial c(\theta_i)}{\partial\theta_i}\right)\frac{\partial\theta_i}{\partial\beta_k}=0 \tag{4.6.16}$$

However, differentiating the LHS of (4.6.15) we have

$$\frac{\partial}{\partial\beta_k}\left\{g\left(-c'(\theta_i)\right)\right\}=g'\left(-c'(\theta_i)\right)\left(-c''(\theta_i)\right)\frac{\partial\theta_i}{\partial\beta_k}=g'(\mu_i)V(\mu_i)\frac{\partial\theta_i}{\partial\beta_k}$$

by (4.6.14), and hence, differentiating the RHS of (4.6.15) and substituting back

$$\frac{\partial\theta_i}{\partial\beta_k}=\frac{1}{g'(\mu_i)V(\mu_i)}\frac{\partial\eta_i}{\partial\beta_k}.$$

In addition,

$$\frac{\partial c(\theta_i)}{\partial\theta_i}=c'(\theta_i)=-\mu_i \qquad \frac{\partial\eta_i}{\partial\beta_k}=x_{ik}$$

so that, finally, (4.6.16) becomes

$$\sum_{i=1}^{n}\frac{w_i}{\phi}\frac{(y_i-\mu_i)x_{ik}}{g'(\mu_i)V(\mu_i)}=0. \tag{4.6.17}$$

Removing the constant ϕ and re-arranging gives the set of score equations

$$\sum_{i=1}^{n}\frac{w_i(y_i-\mu_i)}{V(\mu_i)}\frac{x_{ik}}{g'(\mu_i)}=0 \qquad k=1,...,K \tag{4.6.18}$$

Note that if the canonical link is used

$$\theta_i = x_i^T \beta \Longrightarrow \frac{\partial \theta_i}{\partial \beta_k} = x_{ik}$$

and the equations simplify to

$$\sum_{i=1}^{n} w_i \left(y_i - \mu_i\right) x_{ik} = 0 \qquad \therefore \qquad \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} w_i \mu_i x_{ik} \qquad k = 1, ..., K \tag{4.6.19}$$

which we must solve simultaneously for $\beta_1, ..., \beta_k$ which appear in (4.6.19) as $\mu_i = g^{-1}\left(x_i^T \beta\right)$.

The system (4.6.18) cannot be solved analytically, but can be solved recursively. Let

$$\begin{aligned} Z_i &= \eta_i + \left(Y_i - \mu_i\right) g'\left(\mu_i\right) \qquad (4.6.20) \\ &= x_i^T \beta + \left(Y_i - \mu_i\right) g'\left(\mu_i\right) \end{aligned}$$

then as $E\left[Y_i\right] = \mu_i$ the expected value of Z_i is

$$E\left[Z_i\right] = \eta_i + \left(E\left[Y_i\right] - \mu_i\right) g'\left(\mu_i\right) = \eta_i + 0 \times g'\left(\mu_i\right) = \eta_i.$$

and the variance of Z_i

$$Var\left[Z_i\right] = Var\left[Y_i\right] \left\{g'\left(\mu_i\right)\right\}^2 = \frac{\left\{g'\left(\mu_i\right)\right\}^2 \phi V\left(\mu_i\right)}{w_i}.$$

Hence, from (4.6.20) the $\{Z_i\}$ form a **linear model** in β conditional on $\{\eta_i\}$ and $\{\mu_i\}$, and if observed values $\{z_i\}$ were available, the parameters in $\mu_i = g^{-1}\left(x_i^T \beta\right)$ could be estimated using a **least squares** procedure, that is, by $\widehat{\beta}$ where, (using the **normal equations** from Normal Linear modelling)

$$\widehat{\beta} = \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} z \tag{4.6.21}$$

and where, if $x_i = [x_{i1}, ...x_{iK}]^T$ is the row vector of predictors, $\mathbf{X}$ is the $n \times n$ matrix whose n rows are $x_1, ..., x_n$, and

$$\mathbf{W} = \mathbf{diag}\left(\frac{w_1}{\left\{g'\left(\mu_1\right)\right\}^2 V\left(\mu_1\right)}, ..., \frac{w_n}{\left\{g'\left(\mu_n\right)\right\}^2 V\left(\mu_n\right)}\right)$$

ALGORITHM

The recursive procedure for finding the estimates of β in the GLM then proceeds as follows:

1. **Initialize**: choose any values for $\widehat{\mu}_i^{(0)}$ and $\widehat{\eta}_i^{(0)} = g\left(\widehat{\mu}_i^{(0)}\right)$ for $i = 1, ..., n$

2. For $i = 1, ..., n$ set

$$\widehat{z}_i^{(1)} = \widehat{\eta}_i^{(0)} + \left(y_i - \widehat{\mu}_i^{(0)}\right) g'\left(\widehat{\mu}_i^{(0)}\right)$$

and

$$\widehat{w}_i^{(1)} = \frac{w_i}{\left\{g'\left(\widehat{\mu}_i^{(0)}\right)\right\}^2 V\left(\widehat{\mu}_i^{(0)}\right)} \qquad \Longrightarrow \mathbf{W}^{(1)} = \mathbf{diag}\left(\widehat{w}_1^{(1)}, ..., \widehat{w}_n^{(1)}\right)$$

3. Compute estimates

$$\hat{\beta}^{(1)} = \left(\mathbf{X}^T\mathbf{W}^{(1)}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^{(1)}z^{(1)}$$

4. Define vectors

$$\widehat{\eta}^{(1)} = \mathbf{X}\hat{\beta}^{(1)} \qquad \widehat{\mu}^{(1)} = g^{-1}\left(\widehat{\eta}^{(1)}\right)$$

5. Return to 1 and iterate through 1-5 using $\widehat{\mu}^{(1)}$ and $\widehat{\eta}^{(1)}$ as the updated starting values, eventually to obtain $\widehat{\mu}^{(2)}$ and $\widehat{\eta}^{(2)}$

6. Repeat until $\widehat{\mu}^{(t)}$ and $\widehat{\eta}^{(t)}$ satisfy

$$\left|\widehat{\mu}^{(t)} - \widehat{\mu}^{(t-1)}\right| < \epsilon_\mu \qquad \left|\widehat{\eta}^{(t)} - \widehat{\eta}^{(t-1)}\right| < \epsilon_\eta$$

   for **tolerances** ϵ_μ and ϵ_η.

This process is known as **ITERATIVELY REWEIGHTED LEAST SQUARES**. It can be shown that, as $n \to \infty$,

$$\widehat{\beta} \to N_K\left(\beta, I\left(\beta\right)^{-1}\right) \qquad I\left(\beta\right) = \frac{1}{\phi}\mathbf{X}^T\mathbf{W}\mathbf{X} \qquad [\mathbf{W}]_{ii} = \frac{w_i}{V\left(\mu_i\right)\left\{g'(\mu_i)\right\}^2}$$