3 CATEGORICAL DATA ANALYSIS

Maximum likelihood estimation procedures for loglinear and logistic regression models are discussed in this chapter.

3.1 LOGLINEAR ANALYSIS

3.1.1 The Model

Consider a completely classified contingency table and arrange the observed frequencies into a vector $\mathbf{y}'=(y_1, y_2, y_3, \ldots, y_p)$. The expected cell frequencies are given by $\boldsymbol{\mu}'=(\mu_1, \mu_2, \mu_3, \ldots, \mu_p)$. A Poisson sampling scheme is assumed.
For independent Poisson sampling the joint probability function of Y_i, $i = 1, 2, \ldots, p$ is

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\mu}) &= \prod_{i=1}^{p} \frac{\exp^{-\mu_i}\mu_i^{y_i}}{y_i!} \\ &= \exp\left[\sum y_i \log \mu_i - \sum \mu_i\right] \exp\left[-\sum \log y_i!\right] \end{aligned} \tag{34}$$

which is a member of the exponential family since it has the form

$$p(\mathbf{y}, \boldsymbol{\theta}) = b(\mathbf{y}) \exp\left[\mathbf{y}'\boldsymbol{\theta} - \kappa(\boldsymbol{\theta})\right]$$

with $b(\mathbf{y}) = \exp\left[-\sum \log y_i!\right]$
$\boldsymbol{\theta}$ a 4×1 vector of natural parameters with $\theta_i = \log \mu_i$, that is $\mu_i = \exp(\theta_i)$
$\kappa(\boldsymbol{\theta}) = \sum \mu_i = \sum \exp(\theta_i)$.

The expected value of Y_i is

$$\begin{aligned} E(Y_i) &= \frac{\partial}{\partial \theta_i}\kappa(\boldsymbol{\theta}) \\ &= e^{\theta_i} \\ &= \mu_i \end{aligned}$$

and the covariance of Y_i, Y_j is

$$\begin{aligned} \mathrm{Cov}(Y_i, Y_j) &= \frac{\partial^2}{\partial \theta_i \partial \theta_j}\kappa(\boldsymbol{\theta}) \\ &= \begin{cases} e^{\theta_i} & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Thus $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{Y}) = \mathrm{Diag}(\boldsymbol{\mu})$.

In the case of a 2×2 contingency table with two categorical variables A and B, the model to be fitted, written as a loglinear model is

$$\begin{aligned} \log \mu_1 &= \alpha + \lambda_1^A + \lambda_1^B + \lambda_{11}^{AB} \\ \log \mu_2 &= \alpha + \lambda_1^A - \lambda_1^B - \lambda_{11}^{AB} \\ \log \mu_3 &= \alpha - \lambda_1^A + \lambda_1^B - \lambda_{11}^{AB} \\ \log \mu_4 &= \alpha - \lambda_1^A - \lambda_1^B + \lambda_{11}^{AB} \end{aligned}$$

The generalized linear model is

$$\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}.$$

The three components of the GLM are:

1. The random component $\mathbf{Y}$.

2. The systematic component

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \lambda_1^A \\ \lambda_1^B \\ \lambda_{11}^{AB} \end{pmatrix}$$

where $\mathbf{X}$ is the design matrix and $\boldsymbol{\beta}' = \left(\alpha, \lambda_1^A, \lambda_1^B, \lambda_{11}^{AB}\right)$ the vector with model parameters.

3. The link function is also a canonical link and is given by

$$\eta_i = h(\mu_i) = \log \mu_i = \theta_i = \sum_j \beta_j x_{ij}. \tag{35}$$

3.1.2 Newton-Raphson algorithm for ML estimation

From equation (34) the log-likelihood function for independent Poisson sampling is

$$L(\boldsymbol{\mu}|\mathbf{y}) = \sum_i y_i \log \mu_i - \sum_i \mu_i - \sum_i \log y_i!. \tag{36}$$

In equation (35) $\log \mu_i$ was written as $\log \mu_i = \sum_j \beta_j x_{ij}$. By substituting $\mu_i = \exp\left(\sum_j \beta_j x_{ij}\right)$ into the log-likelihood function in (36), the log-likelihood can be written as a function of the elements of $\boldsymbol{\beta}$. That is

$$L(\boldsymbol{\beta}|\mathbf{y}) = \sum_i y_i \sum_j \beta_j x_{ij} - \sum_i \exp\left(\sum_j \beta_j x_{ij}\right) - \sum_i \log y_i!. \tag{37}$$

The value of $\widehat{\boldsymbol{\beta}}$ that maximizes $L(\boldsymbol{\beta}|\mathbf{y})$ can be found iteratively with the Newton-Raphson algorithm

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \left(\mathbf{H}^{(r)}\right)^{-1} \mathbf{q}^{(r)} \tag{38}$$

where $\boldsymbol{\beta}^{(r)}$ is the rth approximation of $\widehat{\boldsymbol{\beta}}$, $r = 0, 1, 2, \ldots$ and $\mathbf{q}^{(r)}$ and $\mathbf{H}^{(r)}$ are $\mathbf{q}$ and $\mathbf{H}$ evaluated at $\boldsymbol{\beta}^{(r)}$. From Section 2.1, $\mathbf{q}$ is the vector with elements the first order partial derivatives

$$q_k = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_k} = -\sum_i x_{ik} \exp\left(\sum_j \beta_j x_{ij}\right) + \sum_i y_i x_{ik}$$

and $\mathbf{H}$ is the matrix of second order partial derivatives having elements

$$h_{hk} = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_h \partial \beta_k} = -\sum_i x_{ih} x_{ik} \exp\left(\sum_j \beta_j x_{ij}\right) = -\sum_i x_{ih} x_{ik} \mu_i.$$

Hence,

$$\mathbf{q}^{(r)} = \mathbf{X}'\left(\mathbf{y} - \boldsymbol{\mu}^{(r)}\right) \tag{39}$$

$$\mathbf{H}^{(r)} = -\mathbf{X}' \text{diag}\left(\boldsymbol{\mu}^{(r)}\right) \mathbf{X} \tag{40}$$

with $\boldsymbol{\mu}^{(r)} = \exp\left(\mathbf{X}\boldsymbol{\beta}^{(r)}\right)$ the rth approximation of $\widehat{\boldsymbol{\mu}}$, $(r = 0, 1, 2, \ldots)$.
Substituting equations (39) and (40) into equation (38) gives

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \left[\mathbf{X}' \text{diag}\left(\boldsymbol{\mu}^{(r)}\right) \mathbf{X}\right]^{-1} \mathbf{X}'\left(\mathbf{y} - \boldsymbol{\mu}^{(r)}\right). \tag{41}$$

The algorithm requires an initial guess, $\boldsymbol{\beta}^{(0)}$, for the values that maximizes the function $L(\boldsymbol{\beta}|\mathbf{y})$. The ML estimates of the parameters in the saturated model are used as the initial estimates and are given by

$$\boldsymbol{\beta}^{(0)} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \log \mathbf{y}.$$

The asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\text{Cov}\left(\widehat{\boldsymbol{\beta}}\right) = \left[\mathbf{X}'\text{diag}\left(\widehat{\boldsymbol{\mu}}\right)\mathbf{X}\right]^{-1} = -\widehat{\mathbf{H}}^{-1}.$$

A canonical link function was used in the GLM in which case the observed and expected second derivative matrices are identical. Hence, the Fisher scoring and Newton-Raphson algorithms are identical.

3.1.3 Maximum likelihood estimation under constraints

This procedure is also discussed by Crowther and Matthews (1995).
The saturated loglinear model can be written as

$$\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{42}$$

where $\boldsymbol{\mu}' = \left(\mu_1, \mu_2, \mu_3, \ldots, \mu_p\right)$ is the vector with expected cell frequencies for the model, $\mathbf{X} : p \times p$ is the design matrix and $\boldsymbol{\beta} : p \times 1$ is the vector of parameters for the saturated loglinear model. The ML estimate of $\boldsymbol{\beta}$ for the saturated model is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\log \mathbf{y}.$$

For a lower order model certain elements of $\boldsymbol{\beta}$ will be equal to zero.
Let $\mathbf{C}$ be a matrix specifying the elements of $\boldsymbol{\beta}$ which are set equal to zero. The hypothesis that certain elements of $\boldsymbol{\beta}$ are zero, can be written as the constraint

$$\begin{aligned} \mathbf{g}\left(\boldsymbol{\mu}\right) &= \mathbf{C}\boldsymbol{\beta} \\ &= \mathbf{C}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\log \boldsymbol{\mu} \\ &= \mathbf{A}_C'\log \boldsymbol{\mu} \\ &= \mathbf{0}. \end{aligned} \tag{43}$$

The ML estimate of $\boldsymbol{\mu}$ subject to the constraint $\mathbf{g}\left(\boldsymbol{\mu}\right) = \mathbf{A}_C'\log \boldsymbol{\mu} = \mathbf{0}$ is given by

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - \left(\mathbf{G}_\mu \mathbf{V}_\mu\right)'\left(\mathbf{G}_\mathbf{y}\mathbf{V}_\mu\mathbf{G}_\mu'\right)^{-1}\mathbf{g}\left(\mathbf{y}\right) + o\left(\|\mathbf{y} - \boldsymbol{\mu}\|\right)$$

where $\mathbf{G}_\mu = \dfrac{\partial}{\partial \boldsymbol{\mu}}\mathbf{g}\left(\boldsymbol{\mu}\right) = \mathbf{A}_C'\mathbf{D}_\mu^{-1}$ and $\mathbf{V}_\mu = \mathbf{D}_\mu$.
Thus

$$\begin{aligned} \widehat{\boldsymbol{\mu}}_c &= \mathbf{y} - \left(\mathbf{A}_C'\mathbf{D}_\mu^{-1}\mathbf{D}_\mu\right)'\left(\mathbf{A}_C'\mathbf{D}_\mathbf{y}^{-1}\mathbf{D}_\mu\mathbf{D}_\mu^{-1}\mathbf{A}_C\right)^{-1}\mathbf{g}\left(\mathbf{y}\right) + o\left(\|\mathbf{y} - \boldsymbol{\mu}\|\right) \\ &= \mathbf{y} - \mathbf{A}_C\left(\mathbf{A}_C'\mathbf{D}_\mathbf{y}^{-1}\mathbf{A}_C\right)^{-1}\mathbf{g}\left(\mathbf{y}\right) + o\left(\|\mathbf{y} - \boldsymbol{\mu}\|\right). \end{aligned} \tag{44}$$

The ML estimate for $\widehat{\boldsymbol{\mu}}_c$ is obtained by iterating over $\mathbf{y}$ and the asymptotic covariance matrix of $\widehat{\boldsymbol{\mu}}_c$ is

$$\widehat{\boldsymbol{\Sigma}}_c = \mathbf{D}_{\widehat{\boldsymbol{\mu}}_c} - \mathbf{A}_C\left(\mathbf{A}_C'\mathbf{D}_{\widehat{\boldsymbol{\mu}}_c}^{-1}\mathbf{A}_C\right)^{-1}\mathbf{A}_C'.$$

The ML estimate for the vector of cell probabilities is

$$\widehat{\mathbf{p}}_c = \frac{\widehat{\boldsymbol{\mu}}_c}{n}$$

where n is the number of observations.
The ML estimates for the parameters in the loglinear model are given by

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\log \widehat{\boldsymbol{\mu}}_c.$$

The covariance matrix $\widehat{\boldsymbol{\beta}}$ is

$$\text{Cov}\left(\widehat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\text{Cov}\left[\log \widehat{\boldsymbol{\mu}}_c\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

The "delta method" is used to determine the asymptotic covariance matrix

$$\begin{aligned} est\left[\text{Cov}\left(\log\widehat{\boldsymbol{\mu}}_c\right)\right] &= \left(\frac{\partial \log\widehat{\boldsymbol{\mu}}_c}{\partial\widehat{\boldsymbol{\mu}}_c}\right)\widehat{\boldsymbol{\Sigma}}_c\left(\frac{\partial \log\widehat{\boldsymbol{\mu}}_c}{\partial\widehat{\boldsymbol{\mu}}_c}\right)' \\ &= \mathbf{D}_{\widehat{\boldsymbol{\mu}}_c}^{-1}\widehat{\boldsymbol{\Sigma}}_c\mathbf{D}_{\widehat{\boldsymbol{\mu}}_c}^{-1}. \end{aligned}$$

Hence, the estimated covariance matrix for $\widehat{\boldsymbol{\beta}}$ is

$$est\left[\text{Cov}\left(\widehat{\boldsymbol{\beta}}\right)\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left[\mathbf{D}_{\widehat{\boldsymbol{\mu}}_c}^{-1}\widehat{\boldsymbol{\Sigma}}_c\mathbf{D}_{\widehat{\boldsymbol{\mu}}_c}^{-1}\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

EXAMPLE 3.1

Maximum likelihood estimation for a loglinear model.

Pugh (1983) designed a study to examine the disposition of jurors to base their judgments of defendants ("guilty" or "not guilty") on the alleged behavior of a rape victim. Pugh's study varied the degree to which the juror could assign fault to the victim ("low" or "high") and the presentation of the victim as someone with "high moral character", "low moral character" or "neutral". The data are given in Table 3.1.

TABLE 3.1: Data from Pugh (1983).

		Moral (M)		
Verdict (V)	Fault (F)	High	Neutral	Low
Guilty	Low	42	79	32
	High	23	65	17
Not Guilty	Low	4	12	8
	High	11	41	24

The saturated model, $\log\left(\mu_{ijk}\right) = \alpha + \lambda_i^M + \lambda_j^V + \lambda_k^F + \lambda_{ij}^{MV} + \lambda_{ik}^{MF} + \lambda_{jk}^{VF} + \lambda_{ijk}^{MVF}$, can be written as

$$\log\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

$$= \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & -1 & 1 & 0 & -1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & -1 & 1 & -1 & 0 & 1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & -1 & -1 & -1 & 0 & -1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & -1 & 0 & 1 & 0 & -1 & -1 & 0 & -1 \\ 1 & 0 & 1 & -1 & 1 & 0 & -1 & 0 & 1 & -1 & 0 & -1 \\ 1 & 0 & 1 & -1 & -1 & 0 & -1 & 0 & -1 & 1 & 0 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \lambda_1^M \\ \lambda_2^M \\ \lambda_1^V \\ \lambda_1^F \\ \lambda_{11}^{MV} \\ \lambda_{21}^{MV} \\ \lambda_{11}^{MF} \\ \lambda_{21}^{MF} \\ \lambda_{11}^{VF} \\ \lambda_{111}^{MVF} \\ \lambda_{211}^{MVF} \end{pmatrix}$$

Consider in this example the reduced model $\log(\mu_{ijk}) = \alpha + \lambda_i^M + \lambda_j^V + \lambda_k^F + \lambda_{ij}^{MV} + \lambda_{jk}^{VF}$ which contains only the interaction terms between Verdict and Fault and between Verdict and Moral.

Results from the Proc Catmod procedure in SAS

The program and output obtained from the PROC CATMOD procedure in SAS are given in the Appendix. The results are summarized in Table 3.2.

TABLE 3.2: Results from SAS: Proc Catmod.

Maximum Likelihood Estimates		
Variable	Par Estimate	Standard Error
λ_1^M	−0.4221	0.1062
λ_2^M	0.6067	0.0811
λ_1^V	0.5520	0.0734
λ_1^F	−0.1941	0.0666
λ_{11}^{MV}	0.2512	0.1062
λ_{21}^{MV}	0.0178	0.0811
λ_{11}^{VF}	0.3823	0.0666

Model Fitting Information	
Likelihood Ratio	2.81
Pearson Chi-Square	2.80

Obtaining the ML estimates by using the Newton-Raphson algorithm

The ML estimates are obtained iteratively with equation (41),

$$\boldsymbol{\beta}_u^{(r+1)} = \boldsymbol{\beta}_u^{(r)} + \left[\mathbf{X}_u' \text{diag}\left(\boldsymbol{\mu}^{(r)}\right) \mathbf{X}_u\right]^{-1} \mathbf{X}_u' \left(\mathbf{y} - \boldsymbol{\mu}^{(r)}\right)$$

where the matrix $\mathbf{X}_u$ is a submatrix of the design matrix, $\mathbf{X}$, of the saturated model and $\boldsymbol{\beta}_u$ is the parameter vector of the reduced model. The model is

$$\log \boldsymbol{\mu} = \mathbf{X}_u \boldsymbol{\beta}_u = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & -1 & 1 & 0 & -1 \\ 1 & 1 & 0 & -1 & 1 & -1 & 0 & -1 \\ 1 & 1 & 0 & -1 & -1 & -1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & -1 & 0 & 1 & -1 \\ 1 & 0 & 1 & -1 & 1 & 0 & -1 & -1 \\ 1 & 0 & 1 & -1 & -1 & 0 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \lambda_1^M \\ \lambda_2^M \\ \lambda_1^V \\ \lambda_1^F \\ \lambda_{11}^{MV} \\ \lambda_{21}^{MV} \\ \lambda_{11}^{VF} \end{pmatrix}.$$

The ML estimates of the parameters for the saturated model are used as an initial guess of $\widehat{\boldsymbol{\beta}}_u$ and are given by

$$\boldsymbol{\beta}_u^{(0)} = \left(\mathbf{X}_u' \mathbf{X}_u\right)^{-1} \mathbf{X}_u' \log \mathbf{y}.$$

The covariance matrix of $\widehat{\boldsymbol{\beta}}_u$ is

$$\text{Cov}\left(\widehat{\boldsymbol{\beta}}_u\right) = \left[\mathbf{X}_u' \text{diag}\left(\widehat{\boldsymbol{\mu}}\right) \mathbf{X}_u\right]^{-1}.$$

The results obtained are the same as in Table 3.2. The program is given in the Appendix.

Obtaining the ML estimates under constraints

For the model $\log(\mu_{ijk}) = \alpha + \lambda_i^M + \lambda_j^V + \lambda_k^F + \lambda_{ij}^{MV} + \lambda_{jk}^{VF}$, the ML estimate of $\boldsymbol{\mu}$ subject to the constraint

$$\mathbf{g}(\boldsymbol{\mu}) = \mathbf{C}\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \boldsymbol{\beta} = \mathbf{0}$$

can be determined iteratively with equation (44),

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - \mathbf{A}_C \left(\mathbf{A}_C' \mathbf{D}_{\mathbf{y}}^{-1} \mathbf{A}_C\right)^{-1} \mathbf{g}(\mathbf{y}) + o(\|\mathbf{y} - \boldsymbol{\mu}\|)$$

where $\mathbf{A}_C' = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Furthermore

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \log \widehat{\boldsymbol{\mu}}_c$$

and

$$est\left[\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}\right)\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[\mathbf{D}_{\widehat{\boldsymbol{\mu}}_c}^{-1}\widehat{\boldsymbol{\Sigma}}_c\mathbf{D}_{\widehat{\boldsymbol{\mu}}_c}^{-1}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

The Wald statistic is 2.79 and the other results obtained are the same as in Table 3.2. The program is given in the Appendix.

3.2 LOGISTIC REGRESSION

3.2.1 The Model

Let Y_i, $i = 1, 2, \ldots, p$ be independent random variables with $Y_i \sim bi(n_i, \pi_i)$. The frequency distribution for the p independent binomial distributions is given in Table 3.3.

TABLE 3.3: Frequency distribution of p independent binomial distributions.

	Subgroups			
	1	2	$\cdots$	p
Successes	y_1	y_2	$\cdots$	y_p
Failures	$n_1 - y_1$	$n_2 - y_2$	$\cdots$	$n_p - y_p$

Suppose that m covariates, $X_1, X_2, \ldots, X_m$, are observed and that at occasion i, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ and y_i is the number of successes in the n_i trials, $i = 1, 2, \ldots, p$. Let $\boldsymbol{\pi}' = (\pi_1, \pi_2, \ldots, \pi_p)$ be the vector with probabilities of a success within each subgroup and $\mathbf{n}' = (n_1, n_2, \ldots, n_p)$ the vector indicating the number of trials within each subgroup.
The joint probability function of $Y_1, Y_2, \ldots, Y_p$ is

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\pi}) &= \prod_{i=1}^{p} P(Y_i = y_i) \\ &= \prod_{i=1}^{p} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \\ &= \exp\left[\log \prod_{i=1}^{p} \binom{n_i}{y_i} + \log \prod_{i=1}^{p} \pi_i^{y_i} + \log \prod_{i=1}^{p} (1-\pi_i)^{(n_i - y_i)}\right] \\ &= \exp\left[\sum_{i=1}^{p} \log \binom{n_i}{y_i} + \sum_{i=1}^{p} y_i \log \pi_i + \sum_{i=1}^{p} (n_i - y_i) \log(1-\pi_i)\right] \\ &= \exp\left[\sum_{i=1}^{p} \log \binom{n_i}{y_i} + \sum_{i=1}^{p} y_i \log \frac{\pi_i}{1-\pi_i} + \sum_{i=1}^{p} n_i \log(1-\pi_i)\right] \end{aligned} \tag{45}$$

which is a member of the natural exp he form

$$p(\mathbf{y}, \boldsymbol{\theta}) = b(\mathbf{y}) \exp[\mathbf{y}'\boldsymbol{\theta} - \kappa(\boldsymbol{\theta})]$$

where

$b(\mathbf{y}) = \prod_{i=1}^{p} \binom{n_i}{y_i}$

$\boldsymbol{\theta}$ a $p \times 1$ vector with natural parameters $\theta_i = \log \frac{\pi_i}{1-\pi_i}$, that is $\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$

$\kappa(\boldsymbol{\theta}) = -\sum_{i=1}^{p} n_i \log(1-\pi_i) = -\sum_{i=1}^{p} n_i \log\left(\frac{1}{1+e^{\theta_i}}\right) = \sum_{i=1}^{p} n_i \log\left(1+e^{\theta_i}\right).$

For the exponential class

$$\begin{aligned} E(Y_i) &= \frac{\partial}{\partial \theta_i}\kappa(\boldsymbol{\theta}) \\ &= n_i \frac{e^{\theta_i}}{1+e^{\theta_i}} \\ &= n_i \pi_i = \mu_i \end{aligned}$$

and

$$\begin{aligned} \mathrm{Cov}(Y_i, Y_j) &= \frac{\partial^2}{\partial \theta_j \partial \theta_i}\kappa(\boldsymbol{\theta}) \\ &= \begin{cases} n_i\pi_i(1-\pi_i) & \text{if } i = j \\ 0 \quad \textit{otherwise.} \end{cases} \end{aligned} \tag{46}$$

Thus, $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{Y}) = \mathbf{V}_\mu$ =diag$[n_i\pi_i(1-\pi_i)]$.
The logistic regression model is written as $\ell_\mu = \mathbf{X}\boldsymbol{\beta}$ with

$$\ell_{i,\mu} = \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} = \sum_{j=0}^{m} \beta_j x_{ij}.$$

The three components for the GLM are:

- The random component $\mathbf{Y}$, the vector of successes.
- The systematic component which relates the linear predictor to a set of explanatory variables,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{p1} & x_{p2} & \cdots & x_{pm} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}.$$

- The link function which links $\mu_i = E(Y_i)$ to η_i,

$$\eta_i = \log \frac{\pi_i}{1-\pi_i} = \log \frac{n_i\pi_i}{n_i - n_i\pi_i} = \log \frac{\mu_i}{n_i - \mu_i} = h(\mu_i).$$

The function h is a canonical link since $h(\mu_i) = \theta_i = \log \frac{\pi_i}{1-\pi_i}$.

3.2.2 Newton-Raphson algorithm for ML estimation

From equation (45) the log-likelihood function for the logistic regression model is

$$L(\boldsymbol{\pi}|\mathbf{y}) = \sum_{i=1}^{p} \log \binom{n_i}{y_i} + \sum_{i=1}^{p} y_i \log \frac{\pi_i}{1-\pi_i} + \sum_{i=1}^{p} n_i \log(1-\pi_i).$$

Since $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{m} \beta_j x_{ij}$,
and $\log(1-\pi_i) = -\log\left[1+\exp\left(\sum_{j=0}^{m} \beta_j x_{ij}\right)\right]$
the log-likelihood function in terms of $\boldsymbol{\beta}$ is given by

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{p} \log \binom{n_i}{y_i} + \sum_{i=1}^{p} y_i \sum_{j=0}^{m} \beta_j x_{ij} - \sum_{i=1}^{p} n_i \log\left[1+\exp\left(\sum_{j=0}^{m} \beta_j x_{ij}\right)\right].$$

The value $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$ can be determined with the Newton-Raphson algorithm. At step $r+1$ $(r=0,1,2,\ldots)$ in the iterative process the approximation of $\widehat{\boldsymbol{\beta}}$ is given by

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \left(\mathbf{H}^{(r)}\right)^{-1} \mathbf{q}^{(r)} \tag{47}$$

where $\mathbf{q}$ is the vector having elements $\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_k}$, $\mathbf{H}$ is the matrix having elements $\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_h \partial \beta_k}$, and $\mathbf{q}^{(r)}$ and $\mathbf{H}^{(r)}$ are $\mathbf{q}$ and $\mathbf{H}$ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(r)}$.
The elements of $\mathbf{q}^{(r)}$ can be written as

$$\begin{aligned} q_k^{(r)} &= \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_k}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(r)}} \\ &= \sum_{i=1}^{p} y_i x_{ik} + \sum_{i=1}^{p} n_i x_{ik} \left[\frac{\exp\left(\sum_{j=0}^{m} \beta_j^{(r)} x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{m} \beta_j^{(r)} x_{ij}\right)}\right] \\ &= \sum_{i=1}^{p} x_{ik}\left(y_i + n_i \pi_i^{(r)}\right) \end{aligned}$$

and the elements of $\mathbf{H}^{(r)}$ as

$$\begin{aligned} h_{hk}^{(r)} &= \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_h \partial \beta_k}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(r)}} \\ &= -\sum_{i=1}^{p} x_{ih} x_{ik} n_i \frac{\exp\left(\sum_{j=0}^{m} \beta_j^{(r)} x_{ij}\right)}{\left[1+\exp\left(\sum_{j=0}^{m} \beta_j^{(r)} x_{ij}\right)\right]^2} \\ &= -\sum_{i=1}^{p} x_{ih} x_{ik} n_i \pi_i^{(r)}\left(1-\pi_i^{(r)}\right). \end{aligned}$$

Thus

$$\mathbf{q}^{(r)} = \mathbf{X}'\left(\mathbf{y} - \mathbf{n}'\boldsymbol{\pi}^{(r)}\right) \tag{48}$$

and

$$\mathbf{H}^{(r)} = -\mathbf{X}'\mathrm{Diag}\left[n_i \pi_i^{(r)}\left(1-\pi_i^{(r)}\right)\right]\mathbf{X}. \tag{49}$$

Substituting (48) and (49) into (47) gives

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \left\{\mathbf{X}'\text{Diag}\left[n_i\pi_i^{(r)}\left(1-\pi_i^{(r)}\right)\right]\mathbf{X}\right\}^{-1}\mathbf{X}'\left(\mathbf{y}-\mathbf{n}'\boldsymbol{\pi}^{(r)}\right) \tag{50}$$

where

$$\pi_i^{(r)} = \frac{\exp\left(\sum_{j=0}^{m}\beta_j^{(r)}x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{m}\beta_j^{(r)}x_{ij}\right)}. \tag{51}$$

The algorithm requires an initial guess for $\widehat{\boldsymbol{\beta}}$, which is

$$\boldsymbol{\beta}^{(0)} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\ell}$$

where $\boldsymbol{\ell}$ is calculated from the observed data and has elements $\ell_i = \log\frac{\frac{y_i}{n_i}}{1-\frac{y_i}{n_i}}$.
For $r > 0$ the iterative process proceeds by using equations (50) and (51).
The estimated asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is a by-product of the Newton-Raphson algorithm,

$$\text{Cov}\left(\widehat{\boldsymbol{\beta}}\right) = \left\{\mathbf{X}'\text{Diag}\left[n_i\widehat{\pi}_i\left(1-\widehat{\pi}_i\right)\right]\mathbf{X}\right\}^{-1} = -\widehat{\mathbf{H}}^{-1} \tag{52}$$

where $\widehat{\pi}_i$ is the value of $\pi_i^{(r)}$ on convergence.
A canonical link function was used in the GLM in which case the observed and expected second derivative matrices are identical. The Fisher scoring algorithm is identical to the Newton-Raphson algorithm.

3.2.3 Maximum likelihood estimation under constraints

Maximum likelihood estimation for the logistic regression model, using constraints is discussed by Crowther and Matthews (1998).
The logistic regression model can be written as $\boldsymbol{\ell}_\mu = \mathbf{X}\boldsymbol{\beta}$ as discussed in section 3.2.1. The elements of $\boldsymbol{\ell}_\mu$ written as a function of μ_i is

$$\ell_{i,\mu} = \log\frac{\pi_i}{1-\pi_i} = \log\frac{\mu_i}{n_i-\mu_i}.$$

Let $\mathbf{P} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)\mathbf{X}'$ be the projection matrix of the error space. From this the constraint for a logistic regression model as a function of $\boldsymbol{\mu}$ is

$$\mathbf{g}\left(\boldsymbol{\mu}\right) = \mathbf{P}\boldsymbol{\ell}_\mu = \mathbf{P}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}.$$

The ML estimate for $\boldsymbol{\mu}$ is found iteratively with

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - \left(\mathbf{G}_\mu\mathbf{V}_\mu\right)'\left(\mathbf{G}_\mathbf{y}\mathbf{V}_\mu\mathbf{G}_\mu'\right)^{-1}\mathbf{g}\left(\mathbf{y}\right) + o\left(\|\mathbf{y}-\boldsymbol{\mu}\|\right) \tag{53}$$

where $\mathbf{G}_\mu = \frac{\partial\mathbf{g}\left(\boldsymbol{\mu}\right)}{\partial\boldsymbol{\mu}} = \mathbf{P}\mathbf{V}_\mu^{-1}$ since $\frac{\partial\ell_{i,\mu}}{\partial\mu_i} = \frac{1}{n_i\pi_i\left(1-\pi_i\right)}$ and $\mathbf{V}_\mu$ =diag$[n_i\pi_i\left(1-\pi_i\right)]$. Furthermore, $\mathbf{G}_\mathbf{y} = \frac{\partial\mathbf{g}\left(\boldsymbol{\mu}\right)}{\partial\boldsymbol{\mu}}|_{\mu=\mathbf{y}} = \mathbf{P}\mathbf{V}_\mathbf{y}^{-1}$ and $\mathbf{g}\left(\mathbf{y}\right) = \mathbf{P}\boldsymbol{\ell}_\mathbf{y}$ where $\boldsymbol{\ell}_\mathbf{y}$ has elements $\ell_{i,y} = \log\frac{y_i}{n_i-y_i}$. Substituting this into (53) gives

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_c &= \mathbf{y} - \left(\mathbf{P}\mathbf{V}_\mu^{-1}\mathbf{V}_\mu\right)'\left(\mathbf{P}\mathbf{V}_\mathbf{y}^{-1}\mathbf{V}_\mu\mathbf{V}_\mu^{-1}\mathbf{P}\right)^{-1}\mathbf{P}\boldsymbol{\ell}_\mathbf{y} + o\left(\|\mathbf{y}-\boldsymbol{\mu}\|\right)\\ &= \mathbf{y} - \mathbf{P}\left(\mathbf{P}\mathbf{V}_\mathbf{y}^{-1}\mathbf{P}\right)^{-1}\mathbf{P}\boldsymbol{\ell}_\mathbf{y} + o\left(\|\mathbf{y}-\boldsymbol{\mu}\|\right).\end{aligned}$$

Iteration takes place over $\mathbf{y}$.
The asymptotic covariance matrix of $\widehat{\boldsymbol{\mu}}_c$ is

$$\widehat{\boldsymbol{\Sigma}}_c = \mathbf{V}_{\widehat{\boldsymbol{\mu}}_c} - \mathbf{P}\left(\mathbf{P}\mathbf{V}_{\widehat{\boldsymbol{\mu}}_c}^{-1}\mathbf{P}\right)^{-1}\mathbf{P}.$$

The ML estimates for the parameters in the model are given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ell_{\widehat{\mu}_c}$$

where $\ell_{\widehat{\mu}_c}$ is the vector of logits at convergence.
The asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\text{cov}\left(\widehat{\boldsymbol{\beta}}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}\left(\ell_{\widehat{\mu}_c}, \ell'_{\widehat{\mu}_c}\right)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

From the "delta method",

$$\begin{aligned} est\left[\text{cov}\left(\ell_{\widehat{\mu}_c}, \ell'_{\widehat{\mu}_c}\right)\right] &= \left(\frac{\partial \ell_{\widehat{\mu}_c}}{\partial \widehat{\mu}_c}\right)\widehat{\boldsymbol{\Sigma}}_c\left(\frac{\partial \ell_{\widehat{\mu}_c}}{\partial \widehat{\mu}_c}\right)' \\ &= \mathbf{V}_{\widehat{\mu}_c}^{-1}\widehat{\boldsymbol{\Sigma}}_c\mathbf{V}_{\widehat{\mu}_c}^{-1} \end{aligned}$$

and hence, the estimated covariance matrix for $\widehat{\boldsymbol{\beta}}$ is

$$est\left[\text{cov}\left(\widehat{\boldsymbol{\beta}}\right)\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[\mathbf{V}_{\widehat{\mu}_c}^{-1}\widehat{\boldsymbol{\Sigma}}_c\mathbf{V}_{\widehat{\mu}_c}^{-1}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

EXAMPLE 3.2
Maximum likelihood estimation for a logistic regression model with a continuous covariate.

The data in Table 3.4, taken from Agresti (1990), was reported by Cornfield (1962) for a sample of male residents of Framingham, Massachusetts, aged 40-59, classified into 8 subgroups according to blood pressure . During a six-year follow-up period, they were classified according to whether they developed coronary heart disease. This is the response variable. The explanatory variable in the model is the value, x_i, which represents the blood pressure in subgroup i, $i = 1, 2, \ldots, 8$.

TABLE 3.4: Cross-Classification of Framingham Men by Blood Pressure and Heart Disease.

		Heart Disease	
Blood Pressure	x_i	Present (y_i)	Absent ($n_i - y_i$)
< 117	111.5	3	153
$117 - 126$	121.5	17	235
$127 - 136$	131.5	12	272
$137 - 146$	141.5	16	255
$147 - 156$	151.5	12	127
$157 - 166$	161.5	8	77
$167 - 186$	176.5	16	83
> 186	191.5	8	35

The model to be fitted is $\ell_{i,\mu} = \log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_i$ which can be written as

$$\ell_{\boldsymbol{\mu}} = \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & 111.5 \\ 1 & 121.5 \\ \vdots & \vdots \\ 1 & 191.5 \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}. \tag{54}$$

Results from the Proc Logistic and Proc Genmod procedures in SAS
The programs and output obtained from the PROC LOGISTIC and PROC GENMOD procedures in SAS are given in the Appendix. The results are summarized in Table 3.5.

TABLE 3.5: Results from SAS: Proc Logistic and Proc Genmod.

Maximum Likelihood Estimates		
Variable	Parameter Estimate	Standard Error
Intercept	−6.0820	0.7243
Blood Pressure	0.0243	0.00484

Model Fitting Information	
Pearson Chi-Square	6.2899
Deviance	5.9092

Obtaining the ML estimates by using the Newton-Raphson algorithm.
The ML estimate of $\boldsymbol{\beta}$ is found iteratively with equation (50) and the covariance matrix is given by equation (52).
The same results as in Table 3.5 are obtained. The program is given in the Appendix.

Obtaining the ML estimates under constraints
The ML estimate for $\boldsymbol{\mu}$ subject to the constraint $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{P}\ell_{\mu} = \mathbf{PX}\boldsymbol{\beta} = \mathbf{0}$ is found iteratively with the equation

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - \mathbf{P}\left(\mathbf{P}\mathbf{V}_{\mathbf{y}}^{-1}\mathbf{P}\right)^{-1}\mathbf{P}\ell_{\mathbf{y}} + o\left(\|\mathbf{y} - \boldsymbol{\mu}\|\right)$$

where $\ell_{\mathbf{y}} = (\ell_{1,y}, \ell_{2,y}, \ldots, \ell_{p,y})$, $\ell_{i,y} = \log\dfrac{y_i}{n_i - y_i}$ for $i = 1, 2, 3 \ldots p$ and $\mathbf{P} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)\mathbf{X}'$.
Iteration takes place only over $\mathbf{y}$.
The maximum likelihood estimates for the parameters are given by

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\ell_{\widehat{\boldsymbol{\mu}}_c}$$

where $\ell_{\widehat{\boldsymbol{\mu}}_c}$ is the vector of logits at convergence.
The asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}\right) = \left\{\mathbf{X}'\mathrm{Diag}\left[n_i\widehat{\pi}_i\left(1 - \widehat{\pi}_i\right)\right]\mathbf{X}\right\}^{-1}.$$

The same results as in Table 3.5 are obtained. The program is given in the Appendix.

EXAMPLE 3.3

Maximum likelihood estimation for a logistic regression model with a categorical covariate (logit model).

Pugh (1983) designed a study to examine the disposition of jurors to base their judgments of defendants on the alleged behavior of a rape victim. Pugh's study varied the degree to which the juror could assign fault to the victim ("low" or "high"). It also varied the presentation of the victim as someone with "high moral character", "low moral character" or "neutral". The response variable is the judgment of the defendant as "guilty" or "not guilty" by the jurors. The data are given in Table 3.6.

TABLE 3.6: Data from Pugh (1983).

		Moral (M)		
Verdict (V)	Fault (F)	High	Neutral	Low
Guilty	Low	42	79	32
	High	23	65	17
Not Guilty	Low	4	12	8
	High	11	41	24

The model to be fitted is $\ell_{i,\mu} = \log \frac{\pi_i}{1-\pi_i} = \alpha + \lambda_1^M x_{i1} + \lambda_2^M x_{i2} + \lambda_1^F x_{i3}$ where

$x_{i1} = 1$ and $x_{i2} = 0$ if Moral = High,
$x_{i1} = 0$ and $x_{i2} = 1$ if Moral = Neutral,
$x_{i1} = -1$ and $x_{i2} = -1$ if Moral = Low,
$x_{i3} = 1$ if Fault = Low,
$x_{i3} = -1$ if Fault = High.

This model assumes no interaction between moral and fault but it can be extended to include the interaction.

The model can be written as the logit model

$$\ell_\mu = \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 0 & -1 \\ 1 & 0 & 1 & -1 \\ 1 & -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \lambda_1^M \\ \lambda_2^M \\ \lambda_1^F \end{pmatrix}.$$

Programs similar to those in Example 3.2 are given in the Appendix and the results are summarized in Table 3.7.

TABLE 3.7: Results for Example 3.3.

Maximum Likelihood Estimates		
Variable	Parameter Estimate	Standard Error
Intercept	1.0783	0.1469
Moral High	0.4553	0.2226
Moral Neutral	0.1210	0.1717
Fault Low	0.7739	0.1355

Model Fitting Information	
Pearson Chi-Square	0.2552
Deviance	0.2554