5 Generalized linear models

5.1 Definition of a generalized linear model

Definition: A generalized linear model has the following three components:

- **Model matrix:**
$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$
of known constants, with associated parameters $\beta = (\beta_1, \ldots, \beta_p)^T$.
- **Link function:** A link function $g(\cdot)$ which links together the mean
$$\mu_i = \mathrm{E}(Y_i),$$
and the **linear component** $\mathbf{x}_i^T\beta$,
$$g(\mu_i) = \mathbf{x}_i^T\beta.$$
- **Exponential family:** Each response Y_i has a distribution that is from a member of the exponential family with pdf
$$f(y;\theta) = \exp\{yb(\theta) + c(\theta) + d(y)\}.$$

5.1.1 Canonical link functions

Often, the natural parameter $b(\theta)$ in
$$f(y;\theta) = \exp\{yb(\theta) + c(\theta) + d(y)\},$$
is used to link the mean μ_i to the linear component $\eta_i = \mathbf{x}_i^T\beta$:
$$g(\mu_i) = b(\theta_i) = \eta_i = \mathbf{x}_i^T\beta$$
This is known as the **canonical link** function. This may or may not provide a satisfactory model. However, it is often used, at least as a starting point, in data analysis.

Family	Response	Canonical link		Range
$Y_i \sim$ Normal (μ, σ_0^2)	Y_i	$g(\mu) = \mu$	identity link	$-\infty < \mu < \infty$
$Y_i \sim$ Poisson (μ)	Y_i	$g(\mu) = \log\mu$	log link	$\mu > 0$
$Y_i \sim$ Binomial (m_i, π)	Y_i/m_i	$g(\pi) = \log(\frac{\pi}{1-\pi})$	logit link	$0 < \pi < 1$

Note: As we shall see in Section 5.6, for a binomial distribution we model the expected proportion, and denote this by $\pi_i = \mathrm{E}(Y_i/m_i)$.

5.2 Estimation

We use MLE to fit a GLM. Unfortunately, there is usually no explicit solution for the maximum likelihood estimates of the elements of β. Therefore, generally, we need an iterative procedure, i.e. **Fisher's method of scoring**, to determine the MLEs. In fact we can show that this is equivalent to an **iterative weighted least squares** procedure.

As in Question 5 on Problem Sheet 4, the idea of weighted least squares is that if the responses Y_i have non-constant variance then we want to weight the contributions of

$$(Y_i - \underbrace{\mu_i(\beta)}_{\text{fitted values}})^2,$$

in the least squares sum by including weighting factors w_i. Thus, the problem is to minimise

$$\sum w_i (Y_i - \mu_i(\beta))^2,$$

for appropriate weights w_i. This leads to the weighted least squares estimator

$$\widehat{\beta}_{\text{WLS}} = (X^T W X)^{-1} X^T W Y,$$

$$W = \text{diag}(w_i) \quad \text{where} \quad w_i = (\text{var}(Y_i))^{-1},$$

if responses are independent. Unfortunately, W often depends on the coefficients β, and thus cannot be used directly. Also in fitting a GLM we need to include the contribution of the link function.

5.2.1 Fisher's method of scoring: iterative weighted least squares

The MLE of β can be obtained as follows. The log likelihood for independent observations $y_1, \ldots, y_n$ is

$$l(\beta) = \log \left\{ \prod_{i=1}^{n} f(y_i, \theta_i) \right\} = \sum_{i=1}^{n} \{ y_i b(\theta_i) + c(\theta_i) + d(y_i) \}$$

Define $\mu_i = E(Y_i) = -\frac{c'(\theta_i)}{b'(\theta_i)}$ (mean), and $\eta_i = g(\mu_i) = \mathbf{x}_i^T \beta$ (linear component).

$\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ip})$ ith values of explanatory variables (i.e. ith row of X).

To obtain the MLE we require the solution of

$$U_j = \frac{\partial l}{\partial \beta_j} = 0 \qquad (j = 1, \ldots, p) \qquad \text{i.e.} \qquad U = \begin{pmatrix} U_1 \\ \vdots \\ U_p \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Consider the log likelihood for y_i.

$$l_i = \log f(y_i; \theta_i) = y_i b(\theta_i) + c(\theta_i) + d(y_i)$$

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$$

since

$$\begin{aligned} \frac{\partial l_i}{\partial \theta_i} &= y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i) \\ \frac{\partial \mu_i}{\partial \theta_i} &= -\frac{c''(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i) b''(\theta_i)}{b'(\theta_i)^2} = b'(\theta_i) \text{var}(Y_i) \\ \frac{\partial \mu_i}{\partial \beta_j} &= \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \cdot \frac{\partial \mu_i}{\partial \eta_i}. \end{aligned}$$

This leads to

$$U_j = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right).$$

Since solution of $U_j = 0$, $j = 1, \ldots, p$, is often intractable, we use Fisher's method of scoring. This is a modification of the multiparameter Newton-Raphson

$$\beta_r = \beta_{r-1} - \left\{ \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right) \right\}^{-1} U(\beta_{r-1}).$$

Fisher's method of scoring replaces the matrix of second derivatives by its expectation. This gives the iterative scoring formula as

$$I_{r-1}\beta_r = I_{r-1}\beta_{r-1} + U_{r-1}.$$

Therefore, the (j,k)th element of I is

$$\begin{aligned} I_{jk} &= \sum_{i=1}^{n} \text{E} \left(\frac{\partial l_i}{\partial \beta_j} \cdot \frac{\partial l_i}{\partial \beta_k} \right) \\ &= \sum_{i=1}^{n} \text{E} \left\{ \frac{(Y_i - \mu_i)^2}{(\text{var}(Y_i))^2} x_{ij} x_{ik} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right\} \\ &= \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2. \end{aligned}$$

In matrix notation, this is

$$I = X^T W X \text{ with } W = \text{diag}\,(w_{ii}) \quad \text{where } w_{ii} = \frac{1}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2.$$

Therefore, the jth element of $(I\beta + U)$ is

$$\sum_{k=1}^{p} \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \beta_k + \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$$

$$= \sum_{i=1}^{n} \frac{x_{ij}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \left\{ \eta_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} \right\}.$$

Thus, the MLE is determined by the solution of the system of p equations

$$X^T W X \widehat{\beta} = X^T W z, \quad \text{where } z \text{ has elements } z_i = \eta_i + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right),$$

which is of weighted least squares form. However, if W, z depend on β, then we must iterate to obtain MLE of β.

Example

- Fitting Poisson (x_i, Y_i) regression model. Y_i independent $\sim Po(\mu_i = \beta_1 + \beta_2 x_i)$. The model assumes the **identity link**, i.e. $g(\mu_i) = \mu_i$.

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

This model has $\mu_i = E(Y_i) = \text{var}(Y_i)$, and $g(\mu_i) = \mu_i = \eta_i = \beta_1 + \beta_2 x_i$, i.e. $\frac{\partial \mu_i}{\partial \eta_i} = 1$. Thus, $w_{ii} = \frac{1}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{1}{\beta_1 + \beta_2 x_i} = \mu_i^{-1}$ (depends on β). Thus, we have

$$I = X^T W X = \begin{pmatrix} \sum_{i=1}^n \mu_i^{-1} & \sum_{i=1}^n x_i \mu_i^{-1} \\ \sum_{i=1}^n x_i \mu_i^{-1} & \sum_{i=1}^n x_i^2 \mu_i^{-1} \end{pmatrix} \quad \text{and} \quad X^T W z = \begin{pmatrix} \sum_{i=1}^n y_i \mu_i^{-1} \\ \sum_{i=1}^n x_i y_i \mu_i^{-1} \end{pmatrix}.$$

Iterate $\widehat{\beta} = (X^T W X)^{-1} X^T W z$ to find MLE of β.

5.3 Fitting models in R/S-PLUS

R/S-PLUS are powerful **statistical** environments for data analysis, and may be used to fit and analyse GLMs. R has almost identical commands to S-PLUS but is free to download[1].

In R/S-PLUS, if we select a distributional **family** for the response variable, then we are automatically given the **canonical link**. So to use the **canonical link** function we just need to give name of distributional family, e.g. for the Poisson distribution

```
glm (y ~ x, family = poisson)
```

fits the model with **canonical** link $\log \mu_i = \eta_i$ (i.e. g is the log link function). However, to use a **non-canonical** link we need to specify the link argument in family, e.g. for the Poisson distribution

```
glm(y ~ x, family = poisson(identity))
```

fits a model with a **non-canonical** link $\mu_i = \eta_i$ (i.e. g is the identity function).

Example

- Textile data — Poisson regression with identity link.

*****Example — Poisson Regression in R/S-PLUS*****

5.4 Analysis of deviance

Definition of deviance: The **deviance** associated with a model ω is given by

$$D = -2 \log LR = -2 \log \frac{\max L(\text{under model } \omega)}{\max L(\text{under model with } n \text{ parameters, } \Omega)}$$

Use ω to denote the model under consideration, and use Ω to denote the **maximal** or **saturated** model with n parameters. The maximal model Ω has $\widehat{\mu}_i = y_i$, i.e. the model gives a perfect fit since there are n parameters and n observations. The statistic D is the (general) likelihood ratio test statistic. We could also write D in terms of the difference of log likelihoods:

$$D = -2 \{ l(\text{fitted model}, \omega) - l(\text{maximal model or saturated model}, \Omega) \}$$

[1] See http://www.r-project.org/

Example

- Deviance for Poisson family, log link. $Y_1,\ldots,Y_n$ independent with $Y_i \sim Po(\lambda_i)$.

$$l(\beta;y_1,\ldots,y_n) = \sum y_i \log\lambda_i - \sum\lambda_i - \sum\log(y_i!) = \sum y_i\log\mu_i - \sum\mu_i - \sum\log(y_i!)$$

Model Ω: For the saturated (maximal) model we have $\widehat{\mu}_i = y_i$, i.e. best possible fit, and the log likelihood for this maximal model is

$$l(\text{maximal model},\Omega;y_1,\ldots,y_n) = \sum y_i\log y_i - \sum y_i - \sum\log(y_i!)$$

Model ω: For the fitted model, $g(\widehat{\mu}_i) = \mathbf{x}_i^T\widehat{\beta}$ with $g \equiv \log$, as the **log link function** is assumed, i.e. $\widehat{\mu}_i = g^{-1}(\mathbf{x}_i^T\widehat{\beta}) = \exp\{\mathbf{x}_i^T\widehat{\beta}\}$, and

$$l(\widehat{\beta};y_1,\ldots,y_n) = \sum y_i\log\widehat{\mu}_i - \sum\widehat{\mu}_i - \sum\log(y_i!)\ .$$

Thus, the deviance for model ω is

$$D = -2\left\{l(\widehat{\beta}) - l(\text{maximal model},\Omega)\right\} = -2\sum_{i=1}^{n}\left\{y_i\log\frac{\widehat{\mu}_i}{y_i} + (y_i - \widehat{\mu}_i)\right\}.$$

5.4.1 Testing subsets of parameters

Consider two models:

$$\begin{array}{lll} \omega & : & \eta_i = \beta_1 x_1 + \ldots + \beta_q x_q \\ \Omega & : & \eta_i = \beta_1 x_1 + \ldots + \beta_q x_q + \ldots + \beta_p x_p \quad (p > q). \end{array}$$

To test

$$H_0 : \beta_{q+1} = \ldots = \beta_p = 0,$$

we use the distribution of the **change in deviance** under H_0:

$$D_\omega - D_\Omega \sim \chi^2_{p-q} \quad (p > q).$$

This is obtained by applying the LR test of

$$H_0 : \beta_{q+1} = \ldots = \beta_p = 0 \quad \text{against} \quad H_1 : \beta_i \neq 0 \text{ for at least one } q < i \leq p,$$

since

$$D_\omega - D_\Omega = -2\{l(\widehat{\beta}_\omega) - l(\text{maximal model})\} + 2\{l(\widehat{\beta}_\Omega) - l(\text{maximal model})\} = -2\log LR,$$

where

$$LR = \frac{\max L(\text{under restricted model } \omega)}{\max L(\text{under full model } \Omega)}.$$

Example

- Poisson Deviance. Testing constant mean.

  Assume a log link, $\log\lambda_i = \eta_i$ where $\lambda_i = E(Y_i) = \mu_i$

  $$\begin{aligned} \omega &: \eta_i = \alpha \quad \text{(common mean response)} \\ \Omega &: \eta_i = \alpha_i. \end{aligned}$$

  Under ω: $\log\lambda_i = \alpha \qquad \text{or} \qquad \lambda_i = e^{\alpha} = \gamma$ say
  Fit model by MLE, i.e. $\widehat{\gamma} = \frac{\sum y_i}{n} = \bar{y}$, and $\widehat{\mu}_i = \widehat{\gamma} = \bar{y}$. Therefore, the deviance for model ω is

  $$D_\omega = -2\{l(\widehat{\gamma}) - l(\text{maximal model})\} \quad \text{where} \quad l(\widehat{\gamma}) = \sum y_i \log\widehat{\mu}_i - \sum\widehat{\mu}_i.$$

  Under Ω: $\widehat{\mu}_i = y_i$.

  Change in deviance is given by

  $$D_\omega - D_\Omega = -2\left(\sum y_i \log\bar{y} - \sum\bar{y} - \sum y_i\log y_i + \sum y_i\right) = 2\sum y_i \log\frac{y_i}{\bar{y}}.$$

  This may be written in the form of o (observed) and e (expected under H_0):

  $$-2\log LR = 2\sum o\log\left(\frac{o}{e}\right).$$

  Distribution of change in deviance, under ω (H_0), is $D_\omega - D_\Omega \sim \chi^2_{n-1}$.

5.5 Residuals

Ordinary residuals $y_i - \widehat{\mu}_i$ are not used in a GLM (generally) since they have non-constant variance. Two widely used residuals are:

(i) Pearson residuals

$$r_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}},$$

where $V(\mu_i) = \text{var}(Y_i)$ is the variance function in terms of μ_i.

Example

- Poisson family.

  Since $V(\mu_i) = \text{var}(Y_i) = \mu_i$, the ith Pearson residual is given by $r_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}$.

(ii) Deviance residuals

$$d_i = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{\text{deviance associated with } y_i}$$

In R/S-PLUS these may be obtained from a `glm` object using:

1. `residuals(object.glm, type = 'pearson')` for Pearson residuals.
2. `residuals(object.glm, type ='deviance')` for Deviance residuals.

As in regression, residuals may be used to check the data for outliers, and/or model adequacy.

5.6 Logistic models for binomial data

5.6.1 Tolerance distributions: link functions

Suppose that

$$Y_i \sim \text{Binomial}(m_i, \pi_i) \quad (i = 1, \ldots, n),$$

where the probability of 'success' for an observation in the ith group, π_i, is given by

$$g(\pi_i) = \eta_i = \mathbf{x}_i^T \beta$$

and $\mathbf{x}_i^T$ is the value of the ith explanatory variable, i.e. ith row of model matrix X.

The curve for π is of sigmoid form and therefore it is natural to model it by a cdf F — this cdf is called a **tolerance distribution**.

Some possibilities for F are:

	Model	$F(\eta)$	F cdf distribution
(a)	Probit	$\pi = \Phi(\eta)$	N(0,1)
** (b)	Logistic	$\pi = \frac{1}{1+e^{-\eta}}$	logistic
(c)	Complementary log-log	$\pi = 1 - \exp(-\exp(\eta))$	extreme value

In the GLM framework for each of these tolerance distributions we have a corresponding link function:

$$\begin{aligned}
\text{(a)} \quad g(\pi) &= \Phi^{-1}(\pi) = \eta \quad \textbf{Probit link} \\
**\text{(b)} \quad g(\pi) &= \log\left(\frac{\pi}{1-\pi}\right) = \eta \quad \textbf{Logit link} \\
\text{(c)} \quad g(\pi) &= \log(-\log(1-\pi)) = \eta \quad \textbf{Complementary log} - \textbf{log link}
\end{aligned}$$

Link (b) is the canonical link function, and we consider this particular form of the binomial model in the following section.

In R/S-PLUS, to fit a binomial model with link (b) use:

```
glm(cbind(y,m-y) ~ x, family = binomial (logit))
```

`cbind(y,m-y)` is a two column matrix with number of successes in first column and number of failures in second column.

5.6.2 Logistic model: logit link function

Suppose now that $Y_i \sim \text{Binomial}(m_i, \pi_i) \quad (i = 1, \ldots, n)$ where

$$\text{logit}\, \pi_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T \beta$$

This is a GLM since it has three elements:

- Model matrix containing rows $\mathbf{x}_i^T$, and coefficients β.
- The link function is the logit function, i.e. $\text{logit}\, \pi_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$, where $\eta_i = \mathbf{x}_i^T \beta$ is the linear component of the model, and $\pi_i = \text{E}(Y_i/m_i)$. This is the 'systematic part' of the model.

- $Y_i \sim \text{Binomial}(m_i, \pi_i)$ the 'random part' of the model, and binomial is a member of the exponential family.

The log likelihood function is

$$l(\pi_1, \ldots, \pi_n; y_1, \ldots, y_n) = \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) + \text{constant} \right\}$$

where π_i is given by $\text{logit}\, \pi_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T \beta$

The MLEs of π_i, for the saturated model are $p_i = \frac{Y_i}{m_i}$, i.e. the observed proportion of 'success' responses associated with the ith row $\mathbf{x}_i^T$ of $\mathbf{X}$. Thus, for the saturated model $\widehat{\pi}_i = p_i$.

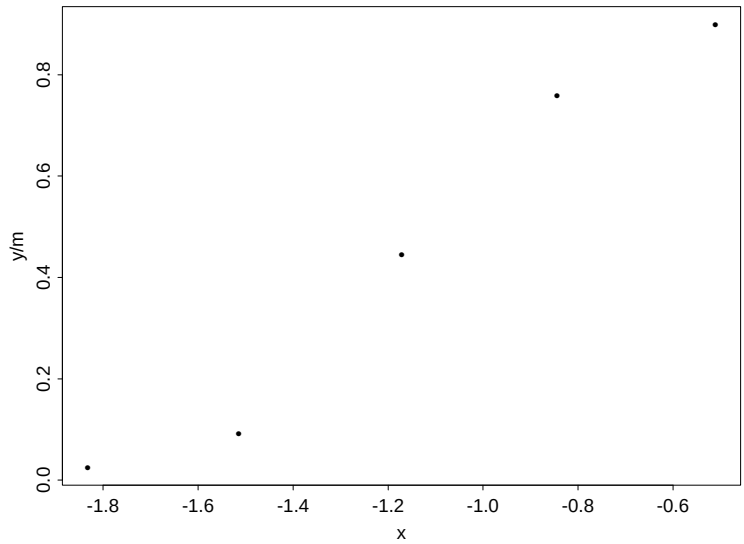
Example

- Dose-response curve. Weevil data set.

  Five doses of an insecticide (*Malathion*) were applied to granary weevils. For each dose (d_i), the number of insects (m_i) receiving that level of dose and the number killed (y_i) were recorded.

Group	Dose	m_i	y_i	p_i
1	0.16	120	3	3/120
2	0.22	120	11	11/120
3	0.31	119	53	53/119
4	0.43	120	91	91/120
5	0.60	119	107	107/119

  Take $x = \log(\text{Dose})$ as the explanatory variable.



$\widehat{\pi}_i$ = estimate of prob $\pi_i = p_i$
Number killed $Y_i \sim \text{Binomial}(m_i, \pi_i)$
where $\pi_i = \Pr$ (insect killed $|x_i$).

5.6.3 Estimation

The binomial model implies that for the ith observation the log likelihood contribution is

$$l(y_i;\pi_i) = y_i\eta_i + m_i\log(1-\pi_i) \quad (\text{“the random part”})$$

where

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = g(\pi_i)$$

if there are m_i trials and y_i successes. The systematic or regression part of the model is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \mathbf{x}_i^T\beta = \sum_{j=1}^{p} x_{ij}\beta_j$$

or $\eta = \mathbf{X}\beta$. This implies that

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\mathbf{x}_i^T\beta}}{1+e^{\mathbf{x}_i^T\beta}}$$

$$\begin{aligned} l(\beta) &= \sum_i \left[y_i(x_{i1}\beta_1 + \cdots + x_{ip}\beta_p) - m_i\log(1+e^{\mathbf{x}_i^T\beta})\right] \\ \frac{dl(\beta)}{d\beta_j} &= \sum_i \left[y_i x_{ij} - m_i\frac{x_{ij}e^{\mathbf{x}_i^T\beta}}{1+e^{\mathbf{x}_i^T\beta}}\right] \\ &= \sum_i [y_i - m_i\pi_i(\beta)]x_{ij} \end{aligned}$$

The ML equations for $\widehat{\beta}$ are

$$\frac{dl(\beta)}{d\beta_j} = 0$$

Use iterative weighted least squares to estimate β by MLE:

$$\widehat{\beta} = (X^TWX)^{-1}X^TWz$$

Here, using the general theory developed above, we have $W = \text{diag}(w_{ii})$ where

$$w_{ii} = \frac{m_i}{\pi_i(1-\pi_i)}\left(\frac{\partial\pi_i}{\partial\eta_i}\right)^2,$$

and $z = (z_1,\ldots,z_n)$ with

$$z_i = \eta_i + \left(\frac{y_i - m_i\pi_i}{m_i}\right)\left(\frac{\partial\eta_i}{\partial\pi_i}\right).$$

Note we are using Y_i/m_i as the response (not Y_i).

The (asymptotic) estimated variance-covariance matrix for the MLEs is given by

$$\text{var}(\widehat{\beta}) = (X^TWX)^{-1}$$

with the (j,k)th element of (X^TWX) given by

$$(X^TWX)_{jk} = \sum_{i=1}^{n} m_i\pi_i(1-\pi_i)x_{ij}x_{ik}$$

5.6.4 Analysis of deviance

Consider the full model Ω $\quad g(\pi) = \mathbf{X}\beta \quad$ with $\widehat{\pi}_\Omega = g^{-1}(\mathbf{X}\widehat{\beta}_\Omega)$

and the reduced model ω $\quad g(\pi) = \mathbf{X}_\omega\beta_\omega \quad$ with $\widehat{\pi}_\omega = g^{-1}(\mathbf{X}_\omega\widehat{\beta}_\omega)$

$$l(\widehat{\pi}_\Omega) = \sum\{y\log\widehat{\pi}_\Omega + (m-y)\log(1-\widehat{\pi}_\Omega)\}$$
$$l(\widehat{\pi}_\omega) = \sum\{y\log\widehat{\pi}_\omega + (m-y)\log(1-\widehat{\pi}_\omega)\}$$

Therefore, the change in deviance (or LRT) statistic is

$$\lambda = -2\log LR = 2[l(\widehat{\pi}_\Omega) - l(\widehat{\pi}_\omega)] = 2\sum\left\{y\log\frac{\widehat{\pi}_\Omega}{\widehat{\pi}_\omega} + (m-y)\log\frac{1-\widehat{\pi}_\Omega}{1-\widehat{\pi}_\omega}\right\}$$

In the special case when the full model is the **saturated** model with the # parameters = # observed values of y. Then clearly $\widehat{\pi}_\Omega = \mathbf{p}$, i.e. the MLE of the true probabilities = the observed proportions. The **deviance** for any reduced model is defined as the -2 log(likelihood ratio) statistic for comparing the reduced model with the saturated model.

$$\text{Deviance} = D(\mathbf{p}, \pi_\omega) = 2\sum\left\{y\log\left(\frac{p}{\widehat{\pi}_\omega}\right) + (m-y)\log\left(\frac{1-p}{1-\widehat{\pi}_\omega}\right)\right\}$$

This can be used directly (as in Poisson case), since it does not involve nuisance parameters, for **analysis of deviance**:

$$D_\omega - D_\Omega \sim \chi^2_{p-q} \qquad (p > q)$$

to test $\beta_{q+1} = \ldots = \beta_p = 0$

$$D_\Omega : \text{deviance for model } \eta_i = \beta_1 x_1 + \ldots + \beta_p x_p$$
$$D_\omega : \text{deviance for model } \eta_i = \beta_1 x_1 + \ldots + \beta_q x_q$$

We can also test for **nonlinearity** using

$$D_\Omega \sim \chi^2_{n-p}$$

D_Ω : deviance of model under consideration.
— cf lack of fit in regression. However, see the comments below on the (asymptotic) distribution of deviance.

Example

- Binomial $Y_i \sim Bin(m_i, \pi_i),\ i = 1, \ldots, n$, with logit link, i.e.

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

In this example we are modelling $\pi_i = P(\text{killed}|x_i)$.

*****Example — R/S-PLUS Logistic Regression*****

ML estimates of (α, β) are: $\widehat{\alpha} = 4.889407$ and $\widehat{\beta} = 4.538052$.

To test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ use analysis of deviance:

$$D_\omega - D_\Omega \sim \chi^2_1$$

D_ω deviance for model logit $(\pi_i) = \alpha$
D_Ω deviance for model logit $(\pi_i) = \alpha + \beta x_i$

Change in deviance is

$$D_\omega - D_\Omega = 341.5 \quad \text{(given in R/S-PLUS output)}$$
$$\chi^2_1(5\%) = 3.84$$
$$\Rightarrow \text{reject } H_0 : \beta = 0$$

Consider the $2 \times n$ table

					Regressors
	#"Successes"	#"Failures"	Total	Proportion	$\mathbf{x}_1 \cdots \mathbf{x}_p$
1	y_1	$m_1 - y_1$	m_1	p_1	
2	y_2	$m_2 - y_2$	m_2	p_2	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n	y_n	$m_n - y_n$	m_n	p_n	

For each of the $2n$ cells we can call the observed frequency o $(= y_i$ or $m_i - y_i)$ and the estimated expected frequency e $(= m_i\widehat{\pi}_i$ or $m_i(1 - \widehat{\pi}_i)$ where $\widehat{\pi}_i$ is a function of $\widehat{\beta})$

Then Deviance can be written

$$D = 2\sum o \log\left(\frac{o}{e}\right)$$

where the summation is now over all $2n$ cells of the table. It can easily be shown that when all m_i are large, this is approximately equal to the X^2 statistic

$$X^2 = \sum \frac{(o-e)^2}{e}$$

Both of these statistics have distributions which are asymptotically (as the $m_i \to \infty$) χ^2_{n-q}. Here $n =$ dimension of the saturated model and $q =$ dimension of the reduced model $=$ rank of the $\mathbf{X}$ matrix.

In the special case when $\mathbf{X}$ consists of the single column $\mathbf{1}$, we have the homogeneity model $\pi = \pi_0\mathbf{1}$, or $\eta = \beta_0\mathbf{1}$ where $\beta_0 = \log\left(\frac{\pi_0}{1-\pi_0}\right)$.

The MLE of π_0 under this reduced model is

$$\widehat{\pi}_0 = \frac{\sum y_i}{\sum m_i} = \frac{y_0}{m_0}$$

and the estimated expected frequencies in the ith row are

$$m_i\widehat{\pi}_0 = \frac{m_i y_0}{m_0} \text{ and } m_i(1 - \widehat{\pi}_0) = \frac{m_i(m_0 - y_0)}{m_0}$$

Then the χ^2 statistic is the familiar contingency table statistic for testing the hypothesis

$$\pi_1 = \pi_2 = \cdots = \pi_n$$

This and the deviance $2\sum o \log\left(\frac{o}{e}\right)$ both $\sim \chi^2_{n-1}$ (approximately if the m_i's are large).

Distribution of the deviance

In general -2 log(likelihood ratio) is approximately χ^2_{p-q} as $n \to \infty$. However, the deviance is defined as the -2 log(likelihood ratio) statistic for the special case of testing the reduced model against the **saturated** model with $p = n$. In considering the asymptotic distribution of the deviance two very different kinds of limit can be considered

1. Keep n fixed and let each $m_i \to \infty$
2. Let $n \to \infty$ with m_i not necessarily large

Under (1) our general result will hold and the deviance $\sim \chi^2_{n-q}$. Under (2) the number of parameters in the saturated model $\to \infty$ as $n \to \infty$ and so general ML theory does not hold: Deviance does not have a χ^2 distribution. As an extreme case, consider all $m_i = 1$. Then it is easy to show that

$$D = -2\sum \{\widehat{\pi}\log\widehat{\pi} + (1-\widehat{\pi})\log(1-\widehat{\pi})\}$$

Since it depends only on the fitted probabilities $\widehat{\pi}$, and not on the differences $y_i - \widehat{\pi}_i$, it is clear that D cannot tell us anything about goodness of fit.

So for logistic regression models tests of goodness of fit using the statistic

$$D = 2\sum o \log\left(\frac{o}{e}\right)$$

or the approximate equivalent

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

and comparing these with χ^2_{n-q} are only valid if the m_i are large.

On the other hand the **change** in deviance between 2 non-saturated models of rank p and q will have a χ^2_{p-q} distribution by the general theorem for likelihood ratio tests. So starting with a given maximal model of rank p say, model selection strategies analogous to those for normal regression models (Backwards Elimination, etc) can be followed, using a sequence of χ^2 tests instead of F tests. However once an acceptable model has been chosen its goodness of fit may have to be assessed using graphical methods for residuals, rather than a single 'omnibus' test of deviance against χ^2_{n-q}.

5.6.5 Residuals for logistic regression

If we fit a model we can calculate the fitted values:

$$\widehat{\pi}_i = g^{-1}(\mathbf{x}_i^T\widehat{\beta}) \quad (i = 1,\ldots,n)$$

From these fitted values we can calculate residuals. Residuals for logistic regression models can be defined in terms of the contribution to either χ^2 or D of the ith row of the $2 \times n$ table of successes and failures.

$$\text{'Pearson residual'} \quad r_i = \frac{y_i - m_i\widehat{\pi}_i}{\sqrt{m_i\widehat{\pi}_i(1-\widehat{\pi}_i)}}$$

$$\text{'Deviance residual'} \quad d_i = \text{sign}(y_i - m_i\widehat{\pi}_i)\sqrt{2\left\{y_i\log\left(\frac{p_i}{\widehat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{1-p_i}{1-\widehat{\pi}_i}\right)\right\}}$$

5.7 Log-linear models for Poisson data

An application of equal importance to logistic regression for binomial random variable is log-linear regression for Poisson random variables. Here we have for a single y

$$l(\mu; y) = y\log(\mu) - \mu - \log(y!)$$

The canonical link is $\theta = \log(\mu)$. Therefore, if we use this in a GLM to link the mean μ_i to the linear component η_i we have the log link

$$\eta = \log(\mu) = \mathbf{X}\beta$$

Note that for the log link function we have fitted values

$$\widehat{\mu_i} = g^{-1}(\mathbf{x}_i^T\widehat{\beta}) = \exp(\mathbf{x}_i^T\widehat{\beta}).$$

5.7.1 Estimation

The method of scoring iterative equation, given above is

$$X^T W X\beta = X^T W\mathbf{z}$$

where $W = \text{diag}(\mu_i)$ because $\text{var}(Y_i) = \mu_i$ for the Poisson distribution and $\mathbf{z}$ has elements

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

5.7.2 Analysis of deviance

We obtained the deviance for Poisson family above as

$$D = -2\left\{l(\widehat{\beta}) - l(\text{maximal model})\right\} = -2\sum_{i=1}^{n}\left\{y_i\log\frac{\widehat{\mu}_i}{y_i} + (y_i - \widehat{\mu}_i)\right\}$$

Example

- Contingency table, r rows and c columns, with the cell entries following a Poisson distribution.

  Model ω: $\eta_i = \mu + \alpha_i + \beta_j$

  Model Ω: $\eta_i = \mu + \alpha_i + \beta_j + \gamma_{ij}$

  To test independence, we test the null hypothesis $H_0 : \gamma_{ij} = 0$.

  This is similar to the Poisson example given above. The test statistic for analysis of deviance is of the form

  $$D = 2\sum o\log(\frac{o}{e})$$

  where o denotes the observed value, and e denotes the expected value under model ω.

  Compare with a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

5.7.3 Residuals for Poisson regression

The Pearson residuals are given by

$$r_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}},$$

where $\widehat{\mu}_i = g^{-1}(\mathbf{x}_i^T \widehat{\beta})$.