Chapter 2

Generalised Linear Models

2.1 Introduction

As an introductory example, suppose we are investigating the efficacy of an insecticide; we expose beetles to various concentrations of carbon disulphide for five hours and see how many were killed. Let $Y_1, \ldots, Y_n$ denote the binary variables $Y_i \in \{0, 1\}$ where 0 denotes that the beetle survived and 1 denotes that it did not survive (the insecticide was successful). Let $z_1, \ldots, z_n$ denote the concentrations of carbon disulphide and $\pi_i = \mathbb{P}(Y_i = 1)$, then we assume that $\pi_i = \pi(z_i)$ (the 'success' probability depends only on the concentration) and that $Y_1, \ldots, Y_n$ are independent. We have:

$$\begin{aligned}
\mathbb{P}((Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)) &= \prod_{j=1}^{n} \pi(z_j)^{y_j} (1 - \pi(z_j))^{1-y_j} \\
&= \exp\left\{ \sum_{j=1}^{n} y_j \log \frac{\pi(z_j)}{1 - \pi(z_j)} + \sum_{j=1}^{n} \log(1 - \pi(z_j)) \right\}.
\end{aligned}$$

Now suppose that the 'success' probability $\pi(z)$ depends on the concentration z through the relationship:

$$\log \frac{\pi(z)}{1 - \pi(z)} = \beta_0 + \beta_1 z.$$

Then

$$\begin{aligned}
\mathbb{P}((Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)) &= \prod_{j=1}^{n} \pi(z_j)^{y_j} (1 - \pi(z_j))^{1-y_j} \\
&= \exp\left\{ \beta_0 \sum_{j=1}^{n} y_j + \beta_1 \sum_{j=1}^{n} y_j z_j + \sum_{j=1}^{n} \log(1 - \pi(z_j)) \right\}
\end{aligned}$$

We can see that this is an exponential family in canonical parameters, with paramters (β_0, β_1), sufficient statistics $(T_1, T_2) = (\sum_{j=1}^n y_j, \sum_{j=1}^n z_j y_j)$ and log partition function

$$A(\beta_0, \beta_1) = \sum_{j=1}^{n} (\beta_0 + \beta_1 z_j - \log(1 + e^{\beta_0 + \beta_1 z_j}))$$

Since the log-partition function is concave, with unique maximum, algorithms for locating $(\widehat{\beta}_0, \widehat{\beta}_1)$, the maximum likelihood estimators are efficient.

2.2 Linear Models: Weighted Least Squares

The algorithm for GLM parameter estimation is the same as that for *weighted* least squares regression, therefore we first describe weighted least squares regression. We then put the GLM problem into the weighted least squares framework.

The *weighted least squares* problem considers the observation vector Y as depending on known covariants X, unknown parameters β and an error term in the form:

$$Y = X\beta + \epsilon$$

where Y is an n-vector of observations, X is an $n \times p$ design matrix, β is a p-vector of unknown parameters and ϵ is an n-vector satisfying $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Cov}(\epsilon) = \sigma^2 W^{-1} = \sigma^2 \mathrm{diag}(\frac{1}{w_1}, \ldots, \frac{1}{w_n})$. With the assumption that $\epsilon \sim N(0, \sigma^2 W^{-1})$,

$$L(\beta) = \frac{(\prod_{j=1}^n w_j)^{1/2}}{(2\pi)^{n/2}\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2}(Y - X\beta)^t W (Y - X\beta) \right\}.$$

Typically, the weights depend on β. The WLS estimator $\widehat{\beta}_{\mathrm{WLS}}$ may therefore be different from the maximum likelihood estimator;

$$\widehat{\beta}_{\mathrm{WLS}} = \mathrm{argmin}(Y - X\beta)^t W (Y - X\beta).$$

2.3 Specification of a Generalised Linear Model

Given $Y = (Y_1, \ldots, Y_n)$, independent observations from the same statistical family (possibly different distributions within the family), a GLM requires

1. A *random component* specifying the distribution from which observations come,
2. A *systematic component*, a parameter η_i which, for observation i is given as a linear combination of the unknown parameters $\beta = (\beta_1, \ldots, \beta_p)^t$;

$$\eta_i = \sum_{j=1}^{p} X_{ij}\beta_j$$

3. A *link*, specifying how η_i relates to $\mu_i = \mathbb{E}[Y_i]$; namely a function g such that $g(\mu_i) = \eta_i$.

Random Component In a generalised linear model, each observation y is generated by a distribution within an exponential family where y is the sufficient statistic. The p.d.f. or p.m.f. is:

$$p(y;\theta,\phi) = e^{c(y,\phi)} \exp\left\{\frac{y\theta - B(\theta)}{a(\phi)}\right\}.$$

Here, we can take θ as the canonical parameter, which means that $\frac{y}{a(\phi)}$ is the sufficient statistic and $A(\theta) = \frac{B(\theta)}{a(\phi)}$ is the log partition function. Recall that, for one-parameter exponential families with canonical co-ordinate η, $\mathbb{E}_\eta[T] = A'(\eta)$ and $\mathrm{Var}_\eta(T) = A''(\eta)$, where T is the sufficient statistic (vector) and $'$ denotes derivative with respect to the canonical parameter η; $A(\eta)$ is the log partition function.

Therefore, with canonical parameter θ:

$$\begin{cases} \mathbb{E}[Y] = a(\phi)A'(\theta) = B'(\theta) \\ \mathrm{Var}(Y) = a^2(\phi)A''(\theta) = a(\phi)B''(\theta). \end{cases}$$

Definition 2.1 (Variance Function)**.** *The* variance function *for a GLM is defined as the function V such that*

$$Var(Y) = a(\phi)V(\mu).$$

From the above, it is clear that such a function exists for families of this type. $V(\mu) = B''(\theta)$; the function $B'(\theta)$ is invertible (just as $\dot{A}(\eta)$ is invertible for a canonical exponential family of full rank) so that $\theta = B'^{-1}(\mu)$ and hence $V(\mu) = B''(B'^{-1}(\mu))$. Note:

$$B'(\theta) = \mu, \qquad B''(\theta) = V(\mu).$$

This is a generalisation of the *Gaussian linear model*; the Gaussian distribution may be written as:

$$p(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} \exp\left\{\frac{y\mu - (\mu^2/2)}{\sigma^2}\right\}.$$

Here (clearly) $\phi = \sigma$, $a(\phi) = \sigma^2$, $B(\mu) = \frac{\mu^2}{2}$, $\mathbb{E}[Y] = \mu$ and $\mathrm{Var}(Y) = \sigma^2$. In this case, the variance function is $V(\mu) \equiv 1$.

X is the *design matrix*, consisting of explanatory variables for the n observations and β is a vector of model parameters. The vector θ is known as the *linear predictor* and the function g is the *link*.

2.4 Parameter Estimation

The aim is to find, at least approximately, $\widehat{\beta}_{ML}$, the maximum likelihood estimator of β. For a single observation y,

$$\log L(\theta; y) = c(y; \phi) + \frac{y\theta}{a(\phi)} - \frac{B(\theta)}{a(\phi)}$$

and for n independent observations

$$\log L(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} c(y_i; \phi) + \frac{1}{a(\phi)} \left(\sum_{i=1}^{n} (y_i\theta_i - B(\theta_i) \right).$$

$\widehat{\beta}_{ML}$ is the solution to the likelihood equation

$$\nabla_\beta \log L(\beta) = 0.$$

For a single observation,

$$\frac{\partial}{\partial \beta_j} \log L(\beta; y) = \frac{\partial}{\partial \theta} \log L \frac{d\theta}{d\mu} \frac{d\mu}{d\eta} \frac{\partial \eta}{\partial \beta_j}.$$

Clearly

$$\frac{\partial}{\partial \theta} \log L = \frac{y - B'(\theta)}{a(\phi)}.$$

Recall that: $B'(\theta) = \mu$ and $B''(\theta) = V$, from which: $\frac{d\mu}{d\theta} = V$. Also, $\frac{d\eta}{d\mu} = g'(\mu)$. Since $\eta = \sum \beta_j x_j$, therefore, for a single observation y with covariates $x_1, \ldots, x_p$:

$$\frac{\partial}{\partial \beta_j} \log L(\beta; y) = \frac{(y - \mu)}{a(\phi)} \frac{1}{V} \frac{1}{g'(\mu)} x_j \tag{2.1}$$

and hence, for n observations:

$$\frac{\partial}{\partial \beta_j} \log L(\beta; y_1, \ldots, y_n) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{a(\phi)} \frac{1}{V(\mu_i)} \frac{1}{g'(\mu_i)} x_{ij}$$

With Equation (2.1) in view, an important quantity for a Generalised Linear Model is the *adjusted dependent variable.*

Definition 2.2 (Adjusted Dependent Variable)**.** *for an observable Y_i, the adjusted dependent variable Z_i is defined as:*

$$Z_i = \eta_i + (Y_i - \mu_i) \frac{d\eta_i}{d\mu_i}(\mu_i).$$

We motivate this later; it will turn out to be a key quantity in the algorithm for finding $\widehat{\beta}_{ML}$, but for now we are interested in its variance. Here Z_i consists of two parts, the *linear systematic component* η_i and the *noise* $(Y_i - \mu_i)\frac{d\eta_i}{d\mu_i}(\mu_i)$. We'll see shortly that finding $\widehat{\beta}_{ML}$ will boil down to weighted least squares regression using the adjusted dependent variables.

The variance may be computed relatively easily:

$$\mathrm{Var}(Z_i) = \mathrm{Var}(Y_i)\left(\frac{d\eta_i}{d\mu_i}(\mu_i)\right)^2 = a(\phi)V(\mu_i)g'(\mu_i)^2 =: \frac{1}{W_i}$$

and the likelihood equation may be expressed:

$$\frac{\partial}{\partial\beta_j}\log L(\beta; y_1, \ldots, y_n) = \sum_{i=1}^{n} W_i g'(\mu_i)(y_i - \mu_i)x_{ij}$$

Let

$$s(\beta) := \nabla \log L(\beta)$$

then $\widehat{\beta}_{ML}$ satisfies $s(\widehat{\beta}_{ML}) = 0$. The parameter estimate is obtained approximately by considering the Taylor expansion:

$$s(\beta) = s(\widehat{\beta}_{ML}) + \nabla\nabla \log L(\beta^*)(\beta - \widehat{\beta}_{ML}) = \nabla\nabla \log L(\beta^*)(\beta - \widehat{\beta}_{ML}).$$

Note:

$$-\mathbb{E}_\beta[\nabla\nabla \log L(\beta; Y)] = I(\beta)$$

the Fisher information. The *Fisher scoring algorithm* is: compute $\widehat{\beta}_{m+1}$ such that:

$$I(\widehat{\beta}_m)(\widehat{\beta}_{m+1} - \widehat{\beta}_m) = s(\widehat{\beta}_m)$$

where $\widehat{\beta}_m$ is the estimate after the mth iteration. We propose an algorithm (which is the standard algorithm for GLM parameter estimation) based on the adjusted dependent variables and weighted least squares and then show that this algorithm is equivalent to the Fisher scoring algorithm.

The Algorithm Let $\widehat{\beta}^m$ denote the estimate of the parameter vector $\beta = (\beta_1, \ldots, \beta_p)^t$ after m iterations and $\widehat{\eta}^m$ the estimate of the linear predictor;

$$\widehat{\eta}^m = X\widehat{\beta}^m.$$

The corresponding fitted value for the vector $\mu = (\mu_1, \ldots, \mu_n)^t$ is: $\widehat{\mu}^m$ where $g(\widehat{\mu}_i^m) = \widehat{\eta}_i^m$. Recall $\eta_i = g(\mu_i)$ for $i = 1, \ldots, n$.

To compute the estimates for iterate $m+1$, the vector of estimate for the *adjusted dependent variables* is computed:

$$z_i^m = \widehat{\eta}_i^m + (y_i - \widehat{\mu}_i^m)\left(\frac{dg}{d\mu}\right)(\widehat{\mu}_i^m).$$

Since $\mathrm{Var}(z_i) = \mathrm{Var}(Y_i)\left(\frac{d\eta_i}{d\mu_i}\right)^2$, the *quadratic weights* are proportional to the inverse of the estimated covariance:

$$\frac{1}{W_i^m} = a_i(\phi)\left(\frac{dg}{d\mu}\right)^2(\widehat{\mu}_i^m)V(\widehat{\mu}_i^m)$$

where V is the *variance function*.

Now regress $z_{m.}$ on the covariates $(x_1, \ldots, x_p)$ in the sense of weighted least squares regression, using weights $(W_1^m, \ldots, W_n^m)$. That is, we use general least squares regression with design matrix X and covariance matrix $\mathrm{diag}(\frac{1}{W_i^m}, \ldots, \frac{1}{W_n^m})$. The parameter estimate $\widehat{\beta}$ is the vector $\widehat{\beta}^{m+1}$, from which $\widehat{\eta}_{m+1}$ is constructed. The algorithm terminates when $\|\widehat{\eta}^{m+1} - \widehat{\eta}^m\|$ is sufficiently small.

Explicitly, let $W^m = \mathrm{diag}(W_1^m, \ldots, W_n^m)$, then

$$\widehat{\beta}^{m+1} = (X^t W^m X)^{-1} X^t W^m z^m,$$

which is the solution to:

$$\widehat{\beta}^{m+1} = \mathrm{argmin}_\beta (z^m - X\beta)^t W^m (z^m - X\beta).$$

By taking derivatives, $\widehat{\beta}^{m+1}$ solves:

$$X^t W^m (z^m - X\widehat{\beta}^{m+1}) = 0.$$

Remarks Firstly, z is a linearised form of the link function applied to the data. To first order (if $y - \mu$ is small - which does not hold for Bernoulli data)

$$g(y) \simeq g(\mu) + (y - \mu)g'(\mu).$$

The covariance matrix of z^m is $\mathrm{diag}(\frac{1}{W_1^m}, \ldots, \frac{1}{W_n^m})$, assuming μ (and hence η) is fixed and known.

The algorithm may be initialised by taking $\widehat{\mu}^0 = y$ (estimating the μs by the data). This can lead to problems with (for example) evaluation of $\log 0$; minor adjustments may be used to compensate for this.

Proof of Correctness We show that the updates thus obtained are the same as those for the *Fisher scoring algorithm*.

With constant dispersion $a_1(\phi) = \ldots = a_n(\phi) = a(\phi)$, the likelihood equation is:

$$s(\beta) =: \nabla_\beta \log L(\beta) = 0 \tag{2.2}$$

where

$$s_j(\beta) = \sum_{i=1}^{n} W_i(y_i - \mu_i)\frac{dg}{d\mu}(\mu_i)x_{ij}. \tag{2.3}$$

Now consider Fisher's scoring method using the gradient vector $s(\beta) = \nabla_\beta \log L(\beta)$, where s denotes the score. Then $s(\widehat{\beta}_{ML}) = 0$. Applying Taylor's expansion and using $\nabla\nabla \log L$ to denote the matrix of second derivatives, then for β close to $\widehat{\beta}_{ML}$,

$$0 = s(\widehat{\beta}_{ML}) \simeq s(\beta) + \nabla\nabla \log L(\beta)(\widehat{\beta}_{ML} - \beta).$$

Let

$$I(\beta) := -\mathbb{E}_\beta[\nabla\nabla \log L(\beta)].$$

This is the *Fisher information matrix*. If the probability distribution is an exponential family in its canonical coordinates (for example: logit or Poisson with log link), then $\nabla\nabla \log L$ is deterministic and equal to $\ddot{A}$. The *Fisher scoring algorithm* for obtaining the maximum likelihood proceeds by computing the update $\widehat{\beta}_{m+1}$ which satisfies:

$$I(\widehat{\beta}_m)(\widehat{\beta}_{m+1} - \widehat{\beta}_m) = s(\widehat{\beta}_m). \tag{2.4}$$

If it is not in canonical co-ordinates, then proceed as follows: from Equation (2.3), using $(uv)' = u'v + uv'$:

$$\begin{aligned} I_{jk} &= -\mathbb{E}\left[\frac{\partial s_j}{\partial \beta_k}\right] \\ &= -\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \mu_i)\frac{\partial}{\partial \beta_k}\left(W_i\frac{dg}{d\mu}(\mu_i)x_{ij}\right) + W_i\frac{dg}{d\mu}(\mu_i)x_{ij}\frac{\partial}{\partial \beta_k}(Y_i - \mu_i)\right]. \end{aligned}$$

The first term vanishes on taking expectations, while using:

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{x_{ij}}{g'(\mu_i)},$$

the second reduces to:

$$\sum_{i=1}^{n} W_i \frac{dg}{d\mu}(\mu_i) x_{ij} \frac{\partial \mu_i}{\partial \beta_k} = \sum_i W_i x_{ij} x_{ik},$$

so that

$$I(\widehat{\beta}^m) = X^t W^m X.$$

Using $\widehat{\eta}^m = X\widehat{\beta}^m$,

$$I(\widehat{\beta}^m)\widehat{\beta}^m = X^t W^m \widehat{\eta}^m.$$

Hence, from (2.3) and (2.4), $\widehat{\beta}^{m+1}$ satisfies:

$$(I(\widehat{\beta}^m)\widehat{\beta}^{m+1})_j = \sum_i W_i^m X_{ij}(\widehat{\eta}_i^m + (y_i - \widehat{\mu}_i^m)\frac{dg}{d\mu}(\widehat{\mu}_i^m)).$$

Therefore, from the definition of z^m,

$$I(\widehat{\beta}^m)\widehat{\beta}^{m+1} = X^t W^m z^m$$

so the update for Fisher's scoring algorithm gives

$$\widehat{\beta}^{m+1} = (X^t W^m X)^{-1} X^t W^m z^m$$

which is the same as the update for the weighted least squares algorithm presented, as required. □

2.5 Goodness of Fit

Goodness of fit may be measured in terms of the *deviance*, which is simply $-2\log\lambda(X)$ where $\lambda(X)$ is the likelihood ratio statistic; the ratio between the *full* model, where $\mu_1, \ldots, \mu_n$ are assumed to be different (no modelling assumptions) and the fitted model, constructed under the assumption that $g(\mu_i) = \sum_{j=1}^{p} X_{ij}\beta_j$, where $p < n$.

Definition 2.3 (Deviance)**.** *The* Deviance *is defined as*

$$D = -2\log\frac{L_{current}}{L_{full}}.$$

In general, given a generalised linear model, we use

$$\begin{aligned} \text{deviance} &= -2\log\left(\frac{\text{max. likelihood for current model}}{\text{max. likelihood for full model}}\right) \\ &= -2\log(\text{likelihood ratio for the current model}) \end{aligned}$$

Bernoulli Data When $Y_1, \ldots, Y_n$ are independent Bernoulli trials, with no further assumptions on $(\mu_1, \ldots, \mu_n) = (\mathbb{E}[Y_1], \ldots, \mathbb{E}[Y_n])$. then $\widehat{\mu}_{ML,i} = Y_i$ and hence

$$L_{\text{full}}(\mu_1, \ldots, \mu_n; Y_1, \ldots, Y_n) = 1.$$

Therefore, in this case:

$$D = -2 \log L_{\text{current}}.$$

The deviance may be thought of as the *asymptotic likelihood ratio statistic*. If H_0 : *current model is correct* holds, then

$$D \sim \chi^2_{p-k}$$

where p is the dimension of the parameter space of the full model and k is the dimension of the parameter space of the current model.

2.6 Binomial Data

For $Y_1, \ldots, Y_n$ Bernoulli trials,

$$p(y; \pi) = \exp\left\{ y \log \frac{\pi}{1-\pi} + \log(1-\pi) \right\} \qquad y \in \{0, 1\}.$$

Here there is no scaling parameter ϕ and $\theta = \log \frac{\pi}{1-\pi}$. In logistic regression, the *canonical link* is used; $\mu = \pi$ and $g(\pi) = \theta = \log \frac{\pi}{1-\pi}$ is used.

Probit and Extreme Value Models

The *Probit* model and the *extreme value* model both have the form: the success probability $\pi(x)$ depends on a vector of covariates x through

$$\pi(x) = F(x^t \beta)$$

for a condinuous c.d.f. F. For the *probit model*, $F = \Phi$, the c.d.f. for $N(0, 1)$ and the model where

$$\Phi^{-1}(\pi(x)) = x^t \beta$$

is known as the *probit model*.

Tolerance Distributions and the Probit Model Binary response models are used in toxicology to describe the effect of a toxic chemical dosage on whether a subject dies. In this application, the concept of a *tolerance distribution* provides justification for the model.

Let x denote the dosage (or, often, log dosage) of a toxic chemical. For a randomly selected subject, let $Y = 1$ if the subject dies. Suppose the subject has tolerance T for the dosage, with $\{Y = 1\}$ equivalent to $\{T \leq x\}$. Let $G(t) = \mathbb{P}(T \leq t)$ denote the c.d.f. of their population distribution. For a fixed dosage x, the probability that a randomly selected subject dies is

$$\mathbb{P}(Y = 1) = \pi(x) = \mathbb{P}(T \leq x) = G(x).$$

In many toxicological experiments, the tolerance distribution for the log dosage is approximately normal with some mean μ and standard deviation σ, so that if G is the cdf of a normal distribution, then

$$\pi(x) = G(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

where Φ is the standard normal cdf. This has the form

$$\pi(x) = F(\alpha + \beta x)$$

with $\alpha = -\mu/\sigma$ and $\beta = 1/\sigma$.

2.6.1 Extreme Value Models

The logit and probit models are both symmetric about 0.5. In other words, for the logit and probit models,

$$\text{link}(\pi) = -\text{link}(1-\pi).$$

Logit and probit models are not appropriate when $\pi(x)$ increases from 0 fairly slowly, but approaches 1 quite suddenly. The response curve is

$$\pi(x) = 1 - \exp\left\{-\exp\left\{\alpha + \beta x\right\}\right\}.$$

Alternatively, when $\pi(x)$ departs from 1 slowly, but approaches 0 sharply, we use

$$\pi(x) = \exp\left\{-\exp\left\{-\frac{(x-a)}{b}\right\}\right\}.$$

This model uses the log-log link

$$\log(-\log(\pi(x))) = \alpha + \beta x.$$

This is a special case of the extreme value (or Gumbel) distribution. That cdf equals

$$G(x) = \exp\left\{-\exp\left\{-\frac{(x-a)}{b}\right\}\right\}$$

for parameters $b > 0$ and $-\infty < a < + < \infty$. It has a mean $a + 0.577b$ and standard deviation $\pi b/\sqrt{6}$.

More generally, for a vector of covariates β, the *extreme value model* is defined as the model which uses the log-log link;

$$\log(-\log(\pi(x))) = x^t\beta.$$

2.6.2 Beetle Mortality Example

The table reports the number of beetles killed after 5 hours exposure to gaseous carbon disulphide at various concentrations. In toxicological experiments, it is usual to measure concentration in terms of log dosage.

			fitted values		
log dose	no. beetles	no. killed	log-log	probit	logit
1.691	59	6	5.7	3.4	3.5
1.724	60	13	11.3	10.7	9.8
1.755	62	18	20.9	23.4	22.4
1.784	56	28	30.3	33.8	33.9
1.811	63	52	47.7	49.6	50.0
1.837	59	53	54.2	53.4	53.3
1.861	62	61	61.1	59.7	59.2
1.884	60	60	59.9	59.2	58.8

For the log-log model, the G^2 goodness of fit statistic is 3.5 based on $df = 6$. By contrast, the logit and probit models behave poorly. The G^2 values are 11.1 for the logit and 10.0 for the probit model.

This poor fit is not surprising given the non-symmetric appearance of the data.

Exercises and Solutions for GLMs

1. (a) Suppose that $(Y_1, \ldots, Y_k) \sim \text{Multnomial}(n; p_1, \ldots, p_k)$. That is:

$$\mathbb{P}((Y_1, \ldots, Y_k) = (y_1, \ldots, y_k)) = \frac{n!}{y_1! \ldots y_k!} p_1^{y_1} \ldots p_k^{y_k} \qquad \sum_{i=1}^{k} p_i = 1, \qquad \sum_{i=1}^{k} y_i = n.$$

Consider testing

$$H_0 : \log p_i = \mu + \beta x_i \qquad 1 \leq i \leq k$$

where $(x_i)_{i=1}^k$ is given and μ is such that $\sum_{i=1}^k p_i = 1$. That is;

$$\mu = -\log \sum_{i=1}^{k} e^{\beta x_i}.$$

show that $\widehat{\beta}$ is the solution of:

$$\sum_{i=1}^{k} y_i x_i = \frac{n \sum_{i=1}^k x_i e^{x_i \beta}}{\sum_{i=1}^k e^{x_i \beta}}.$$

(b) Now suppose instead that $Y_1, \ldots, Y_k$ are independent, satisfying $Y_i \sim \text{Poiss}(\mu_i)$ and consider testing the null hypothesis:

$$HP_0 : \log \mu_i = \alpha + \beta x_i \qquad 1 \leq i \leq k.$$

Find $\log L(\beta, \alpha)$, the log likelihood function and hence show that the equation for $\widehat{\beta}_{ML}$ is the same as the equation for multinomial sampling from the previous part of the exercise. Show also that

$$\sum_{i=1}^{k} \mu_i(\widehat{\beta}, \widehat{\alpha}) = \sum_{i=1}^{k} y_i$$

where $\mu_i(\beta, \alpha) := e^{\alpha + \beta x_i}$.

2. A client gives you data for binary regression consisting of the observations (y, x) which are:

$$(0, x_1), (1, x_1), (0, x_1), (1, x_2), (1, x_2)$$

$$(1, x_2), (0, x_2), (1, x_3), (1, x_3), (0, x_3)$$

$$(1, x_4), (1, x_4), (1, x_4), (1, x_4), (0, x_4).$$

y is the value of the response (1 or 0) while x is the value of the regressor (or explanatory) variable. There are 15 observations. She asks you to fit the model:

$$\log \frac{p(x)}{1-p(x)} = \alpha + \beta x$$

where $p(x)$ is the 'success' probability when the value of the regressor is x and to use appropriate differences in the deviances to test $H_0 : \beta = 0$.

You observe that the data can be considered as four independent binomial samples, which can be encoded $(1, 3, x_1)$, $(3, 4, x_2)$, $2, 3, x_3)$, $(4, 5, x_4)$.

Do you get a different answer if you consider it in this way? Write out the likelihood in both cases; 15 independent Bernoulli trials and 4 independent Binomial variables.

3. The observed waiting times $t_1, \ldots, t_n$ are from independent random variables $T_1, \ldots, T_n$ with density from the family:

$$p(t) = \frac{\lambda^\nu t^{\nu-1}}{\Gamma(\nu)} e^{-\lambda t} \mathbf{1}_{[0,+\infty)}(t)$$

For T_i, the parameters are (ν, λ_i). The parameter ν is the same for each i and is fixed and known, while

$$\lambda_i = \alpha + \beta x_i.$$

   (a) Compute the sufficient statistics for (α, β).

   (b) Suppose that $t_1, \ldots, t_m$ are the times between consecutive earthquakes in Mexico city, while $t_{m+1}, \ldots, t_n$ are the times between consecutive earthquakes in Turkey. Let $x_1 = \ldots = x_m = 0$ and $x_{m+1} = \ldots = x_n = 1$. Derive the equations that $(\widehat{\alpha}_{ML}, \widehat{\beta}_{ML})$ must solve and derive the asymptotic distribution of $\widehat{\beta}_{ML}$ as $m \to +\infty$, $n - m \to +\infty$.

4. Suppose $U_i \sim \text{Binomial}(n_i, p_i)$ for $1 \leq i \leq k$ are independent and

$$g(p_i) = \alpha + \beta x_i.$$

   (a) Compute

$$\frac{\partial}{\partial \alpha} \log L(\alpha, \beta). \qquad \frac{\partial}{\partial \beta} \log L(\alpha, \beta)$$

   where L denotes the likelihood.

   (b) What are the sufficient statistics for (α, β) if:

      i. $g(p) = \log \frac{p}{1-p}$

ii. $g(p) = \log(-\log(1-p))$.

5. If asymptotic normality holds, then an asymptotic confidence interval for β can be derived from

$$(\widehat{\beta} - \beta)^t \left(\mathrm{Cov}(\widehat{\beta})\right)^{-1} (\widehat{\beta} - \beta) \sim \chi^2_p$$

where β is a p-vector.

Set a confidence level of α. Show that if $p = 2$ the resulting equation is an ellipse. What is the equation if $\frac{\partial^2}{\partial\beta_1\partial\beta_2} \log L(\beta) = 0$?

Why is it useful to know that $\int_c^\infty e^{-x}dx = e^{-c}$?

6. Let $y_1, \ldots, y_n$ be independent observations, where the p.d.f. for y_i is from a GLM family:

$$p(y; \theta, \psi) = e^{c(y,\phi)} \exp\left\{\frac{y\theta - B(\theta)}{a(\phi)}\right\}.$$

Suppose $g(\mu_i) = \beta^t x_i$ as usual. Find

$$\mathbb{E}\left[\frac{\partial^2}{\partial\phi\partial\beta} \log L(\phi, \beta; Y_1, \ldots, Y_n)\right]$$

and, from the result, conclude that $\widehat{\beta}$ and $\widehat{\phi}$ are asymptotically uncorrelated (hence asymptotically independent if asymptotic normality holds).

Answers

1. (a)

$$\log L(\beta; y_1, \ldots, y_k) = \text{const} + \sum_{j=1}^{k} y_j \log p_j = \text{const} + \mu \sum y_j + \beta \sum x_j y_j$$

so

$$\frac{d}{d\beta} \log L(\beta) = \frac{d\mu}{d\beta} \sum y_j + \sum x_j y_j$$

where

$$\frac{d\mu}{d\beta} = -\frac{\sum_j x_j e^{\beta x_j}}{\sum e^{\beta x_j}}.$$

Since $\sum_j y_j = n$, hence $\widehat{\beta}$ solves:

$$\sum y_j x_j = \frac{n \sum x_j e^{\beta x_j}}{\sum e^{\beta x_j}}.$$

(b)

$$L(\alpha, \beta) = \frac{1}{\prod_{j=1}^{k} y_j!} \exp\{\alpha \sum y_j + \beta \sum x_j y_j\} \exp\{-e^{\alpha} \sum_j e^{\beta x_j}\}$$

$$\begin{cases} \frac{\partial}{\partial \beta} \log L = \sum x_j y_j - e^{\alpha} \sum_j x_j e^{\beta x_j} \\ \frac{\partial}{\partial \alpha} \log L = \sum_j y_j - e^{\alpha} \sum_j e^{\beta x_j} \end{cases}$$

Set this equal to 0 and let $n := \sum_j y_j$ then:

$$e^{\widehat{\alpha}} \sum_j e^{\widehat{\beta} x_j} = n$$

so that $\widehat{\beta}$ satisfies equation above.

PRACTICAL USE FOR THIS: there is no need for separate software for 'multinomial' distribution; simply use the Poisson software.

$$\widehat{\mu}_i = e^{\widehat{\alpha}} e^{\widehat{\beta} x_i}$$

so that

$$\sum_i \widehat{\mu}_i = e^{\widehat{\alpha}} \sum_i e^{\widehat{\beta} x_i} = \sum_i y_i$$

from above.

2. Let Y_{ij} denote the jth observation for covariate value x_i. Assume these are independent with distribution $\text{Be}(p(x_i))$. The data log likelihood is:

$$\log L(\alpha, \beta) = \sum_i \sum_j y_{ij} \log p(x_i) + (1 - y_{ij}) \log(1 - p(x_i)).$$

Let $y_{i+} = \sum_j y_{ij}$ and n_i the number of trials with covariate value x_i, then

$$\log L(\alpha, \beta) = \sum_i y_{i+} \log p(x_i) + (n_i - y_{i+}) \log(1 - p(x_i))$$

so the data log likelihood is clearly the same (up to a constant) irrespective of whether it is considered as binomials or bernoullis.

For the deviance: for the 'full' model, no assumptions are made about the relations between the $\pi_i =: p(x_i)$; they are estimated from the data $\widehat{\pi}_1 = \frac{1}{3}$, $\widehat{\pi}_2 = \frac{3}{4}$, $\widehat{\pi}_3 = \frac{2}{3}$ and $\widehat{\pi}_4 = \frac{4}{5}$. This is the same irrespective of whether Bernoullis or Binomials are considered. The maximum likelihood estimates for $\widehat{\alpha}$ and $\widehat{\beta}$ are also the same; the deviance $D = -2 \log \frac{L_{\text{max}}}{L_{\text{full}}}$ would have asymptotic distribution $\chi^2_{3-2} = \chi^2_1$ if the sample sizes were considerably larger.

3. (a)

$$\log L(\nu, \alpha, \beta) = -n \log \Gamma(\nu) + \nu \sum_{j=1}^{n} \log(\alpha + \beta x_i) + (\nu - 1) \sum_{j=1}^{n} \log t_i - \alpha \sum_i t_i - \beta \sum t_j x_j$$

The sufficient statistic for (ν, α, β) is therefore: $(\sum_j \log t_j, \sum_j t_j, \sum_j t_j x_j)$: for (α, β) take: $(\sum_j t_j, \sum_j x_j t_j)$.

(b) To estimate (α, β) (ν known), maximise:

$$\nu \sum \log(\alpha + \beta x_j) - \alpha \sum t_i - \beta \sum t_i x_i$$

Let $T_1 = \sum_{j=1}^{m} t_j$ and $T_2 = \sum_{j=m+1}^{n} t_j$ then the expression to be maximised may be expressed as:

$$\nu m \log \alpha + \nu(n - m) \log(\alpha + \beta) - \alpha T_1 - (\alpha + \beta) T_2$$

Taking derivatives gives:

$$\begin{cases} \frac{\nu m}{\alpha} + \frac{\nu(n-m)}{\alpha+\beta} - (T_1 + T_2) = 0 \\ \frac{\nu(n-m)}{\alpha+\beta} - T_2 = 0 \end{cases}$$

so

$$\widehat{\alpha} = \frac{\nu m}{T_1} \qquad \widehat{\beta} = \frac{\nu(n-m)}{T_2} - \frac{\nu m}{T_1}.$$

For asymptotic properties of estimators: we have an exponential family in canonical parametrisation $\binom{\widehat{\alpha}}{\widehat{\beta}}$ is asymptotically normal with mean $\binom{\alpha}{\beta}$ and covariance matrix the inverse of the Fisher information matrix.

$$-\frac{\partial^2}{\partial\alpha^2}\log L = \frac{\nu m}{\alpha^2} + \frac{\nu(n-m)}{(\alpha+\beta)^2}$$
$$-\frac{\partial^2}{\partial\beta^2}\log L = -\frac{\partial^2}{\partial\alpha\partial\beta}\log L = \frac{\nu(n-m)}{(\alpha+\beta)^2}$$

so that

$$\mathrm{Cov}(\widehat{\alpha},\widehat{\beta}) = \frac{\alpha^2}{m\nu}\begin{pmatrix} 1 & -1 \\ -1 & 1+\frac{m}{n-m}(1+\frac{\beta}{\alpha})^2 \end{pmatrix}$$

4. (a)

$$\frac{\partial}{\partial\alpha}\log L = \sum_i \frac{u_i}{p_i(1-p_i)g'(p_i)} - \sum_i \frac{n_i}{(1-p_i)g'(p_i)}$$

$$\frac{\partial}{\partial\beta}\log L = \sum_i \frac{u_i x_i}{p_i(1-p_i)g'(p_i)} - \sum_i \frac{n_i x_i}{(1-p_i)g'(p_i)}$$

(b) For $g(p) = \log\frac{p}{1-p}$ $g'(p) = \frac{1}{p(1-p)}$ hence

$$\frac{\partial}{\partial\alpha}\log L = \sum_i u_i - \sum_i n_i p_i \qquad \frac{\partial}{\partial\beta}\log L = \sum_i x_i u_i - \sum_i n_i p_i$$

The sufficient statistic is therefore $(\sum u_i, \sum u_i x_i)$ by the factorisation theorem.

For $g(p) = -\log(-\log(p))$, $g'(p) = \frac{1}{p\log p}$ and the expressions are quite horrible;

$$\frac{\partial}{\partial\alpha}\log L = \sum_i u_i\frac{\log p_i}{1-p_i} - \sum_i n_i\frac{p_i\log p_i}{(1-p_i)}$$

$$\frac{\partial}{\partial\beta}\log L = \sum_i x_i u_i\frac{\log p_i}{1-p_i} - \sum_i n_i x_i\frac{p_i\log p_i}{1-p_i}$$

that there is no reduction; the sufficient statistic is: $(u_1,\ldots,u_k)$.

5. The covariance matrix is positive definite, hence this is the equation for an ellipse.

For $p=2$, use $\chi^2_{2,\alpha}$ to denote value such that $\mathbb{P}(V \geq \chi^2_{2,\alpha}) = \alpha$, then equation is:

$$(\beta_1 - \widehat{\beta}_1)^2 I_{11}(\widehat{\beta}) + (\beta_2 - \widehat{\beta}_2)^2 I_{22}(\widehat{\beta}) = \chi^2_{2,\alpha}$$

where I (the information matrix) is a diagonal matrix.

The remark is useful because $\chi^2_2 = \text{Exp}(\frac{1}{2})$.

6.
$$\frac{\partial^2}{\partial\phi\partial\beta}\log L = -\frac{a'(\phi)}{a^2(\phi)}\sum_{j=1}^{n}\frac{(y_i-\mu_i)}{V(\mu)g'(\mu)}x_j$$

which clearly has mean zero. When asymptotic normality holds, $\binom{\widehat{\beta}}{\widehat{\phi}} \sim N\left(\binom{\alpha}{\phi}, I^{-1}(\beta,\phi)\right)$. Now, if β is a p-vector, then $I(\beta,\phi)$ is a $p+1\times p+1$ matrix with entries $I_{i,p+1}=0$ for $i=1,\ldots,p$ and hence $I_{p+1,i}=0$ for $i=1,\ldots,p$. For matrices of the form $A=\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22}\end{pmatrix}$, $A^{-1}=\begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1}\end{pmatrix}$, hence $I^{-1}_{i,p+1}(\beta,\phi)=0$ for all $i=1,\ldots,p$ $I_{p+1.i}(\beta,\phi)=0$ for $i=1,\ldots,p$ so that $\text{Cov}(\widehat{\phi},\widehat{\beta})=0$. . Hence asymptotic normality gives asymptotic independence.