Chapter 29

MSMYS4 Generalised Linear Modelling

(29.1) Generalised Linear Regression

(29.1.1) Linear Regression

The familiar linear regression model is of the form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon \tag{1}$$

where $\operatorname{var} \varepsilon = \sigma^2$. Here y is the 'response' variable and the X_is are covariates which are functions of the (independent) explanatory variables. (Unfortunately the explanatory variables also get called x.) Observe that the parameters (the βs) are combined in a linear way and that the model is additive in the sense it has a linear part plus an error part. These desirable features will be retained in the generalised linear model.

It is usual to assume that y is Normally distributed and that the parameters **fi** have Normal sampling distributions. Clearly this covers very few cases and the generalised linear model sets out to correct this.

Take for example binary data coming from a Bernoulli distribution with parameter p. Since $0 \leqslant p \leqslant 1$ it would be foolish to assume the sampling distribution of p to be Normal. It will be shown that a suitable way to estimate p is

$$p = \frac{1 + \exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}\right)}{\exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}\right)}$$

Although this clearly has the required properties it does seem rather unwieldy. Observe that re-arranging

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

Similarly for Poisson data with distributional parameter λ it is required that $\lambda > 0$ and hence one could take

$$ln\lambda = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

(29.1.2) Notes On Modelling

Once a model has been fitted it is of interest as to how well it fits the data. Classically the quantity

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

is used, and with good reason. This calculation suggests that measurements of the ys are on the same physical scale, and that measurements are independent. It also suggests that deviations do not depend on y, so are independent on the mean. However, in the generalised case these assumptions are not always true, and other measures such as

$$\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} \quad \text{or} \quad \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i)$$

may be more appropriate. In particular the first of these is suitable for a Poisson variation.

The likelihood function plays an important rôle in assessing the plausibility of values a parameter might take. The probability density function is treat not as a function of x, the independent variable, but of the parameter μ, say. Taking the logarithm of the likelihood function in the Normal case gives a multiple of equation (2). The data values which maximise the likelihood are of course what one might expect to be observed, and information about the parameters is greatest where the slope of the likelihood surface is greatest. Hence a well designed experiment, looking for certain data by appropriate choice if independent variable, can produce better estimates for the parameters.

Trivially, a model with as many parameters as the number of data observed will fit the data perfectly. However, it does not summarise the data and may be exceptionally poor at predicting other data. A good model uses as few parameters as possible to create an acceptable fit to the data.

(29.1.3) The Exponential Family Of Distributions

The generalised linear model is concerned with finding models for data through estimating a single parameter. The data must have an distribution belonging to the exponential family

$$\ln f(y, \theta) = \frac{y\theta - b(\theta)}{\phi} - c(y, \phi) \tag{3}$$

where b and c are specified functions. In equation (3) θ is the parameter, while ϕ is a constant called the scale parameter or dispersion parameter—it roughly corresponds to variance. Most common distributions belong to the exponential family. In the Poisson case

$$\ln f(y, \theta) = \frac{y \ln \lambda - \lambda}{1} - \ln y!$$

so $\theta = \ln \lambda$, $\phi = 1$, $b(\theta) = e^{\theta}$ and $c(y, \phi) = y!$. For the binomial distribution

$$f(y, p) = \binom{n}{y} p^y (1-p)^{n-y}$$

hence taking the natural logarithm

$$\ln f(y, p) = y \ln\left(\frac{p}{1-p}\right) + n \ln(1-p) + \ln \binom{n}{y} \tag{4}$$

so $\theta = \ln\left(\frac{p}{1-p}\right)$ so that $\phi = 1$. Now,

$$\begin{aligned}
\theta &= \ln\left(\frac{p}{1-p}\right) \\
(1-p)e^{\theta} &= p \\
e^{\theta} &= p\left(1+e^{\theta}\right) \\
p &= \frac{e^{\theta}}{1+e^{\theta}} \\
1-p &= \frac{1}{1+e^{\theta}}
\end{aligned}$$

Hence substituting in equation (4) gives

$$\ln f(y,p) = y\theta - n\ln\left(\frac{1}{1+e^{\theta}}\right) + \binom{n}{y}$$

from this it is now clear that $b(\theta) = -n\ln\left(\frac{1}{1+e^{\theta}}\right)$.

The Mean Of An Exponential Family Distribution

An obvious question to ask is to the mean (and in due course the variance) of an exponential family distribution. Let f be the probability density function for an exponential family distribution. Then

$$\begin{aligned}
\int_{-\infty}^{\infty} f(y,\theta)\,\mathrm{d}y &= 1 \quad \text{the limits are independent of } \theta \text{ hence differentiating,} \\
\int_{-\infty}^{\infty} \frac{\partial f}{\partial\theta}\,\mathrm{d}y &= 0 \\
\int_{-\infty}^{\infty} \frac{\partial \ln f}{\partial\theta} f(y,\theta)\,\mathrm{d}y &= 0 \quad \text{because by the chain rule } \frac{\partial \ln f}{\partial\theta} = \frac{\mathrm{d}\ln f}{\mathrm{d}f}\frac{\partial f}{\partial\theta} \qquad (5) \\
\int_{-\infty}^{\infty} f(y,\theta)\frac{\partial}{\partial\theta}\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)\,\mathrm{d}y &= 0 \\
\int_{-\infty}^{\infty} (y-b'(\theta))f(y,\theta) &= 0 \\
\mathbb{E}\,Y &= b'(\theta) \qquad (6)
\end{aligned}$$

The Variance Of An Exponential Family Distribution

Differentiating equation (5) with respect to θ again gives

$$\begin{aligned}
\int_{-\infty}^{\infty} f(y,\theta)\frac{\partial^2 \ln f}{\partial\theta^2} + \frac{\partial \ln f}{\partial\theta}\frac{\partial f}{\partial\theta}\,\mathrm{d}y &= 0 \\
\int_{-\infty}^{\infty} f(y,\theta)\left(\frac{\partial^2 \ln y}{\partial\theta^2} + \left(\frac{\partial \ln y}{\partial\theta}\right)^2\right)\,\mathrm{d}y &= 0 \\
\frac{-b''(\theta)}{\phi} - \frac{\operatorname{var} Y}{\phi^2} &= 0 \\
\operatorname{var} Y &= \phi b''(\theta)
\end{aligned}$$

(29.1.4) Modelling

Consider a random variable Y which is observed at different levels of explanatory random variables $X_1, X_2, \ldots, X_c$. The observed data, y_j say, has associated explanatory variables observed to be $x_{j1}, x_{j2}, \ldots, x_{jc}$ which may be expressed in vector form as $\mathbf{x}_j$.

Suppose Y has an exponential family distribution so that

$$y_i = \exp\left(\frac{y_i\theta_i - b_i(\theta_i)}{\phi} + c_i(y_i, \phi)\right)$$

Note that b is dependent on i; take for example the Binomial case where $\mathbb{E}\,Y = np$. The joint probability density function for the observed data is therefore

$$f(\mathbf{y}, `) = \exp\left(\frac{1}{\phi}\sum_{i=1}^{n}\left(y_i\theta_i - b_i(\theta_i)\right) + \sum_{i=1}^{n} c_i(y_i, \phi)\right)$$

The model would, essentially, be complete by specifying values for ` but there are far too many θ parameters to make a sensible model. A distinction is now made between the basic explanatory variables and the covariates. Explanatory variables are actual observations whereas the covariates are combinations of these, say $x_1 = X_1X_2$ etc. (and in due course other things). Unfortunately the same letter, x is used to refer to them both.

The mess is resolved as follows. Each of the n y_i has associated with it a vector of p covariates

$$\mathbf{x}_i = \begin{pmatrix} 1 & x_{i1} & x_{i2} & \ldots & x_{i(p-1)} \end{pmatrix}$$

For each covariate introduce a parameter β so that $\theta_i = k(\mathbf{x}_i\mathbf{fi})$ so there are now only p parameters.

Non-Normal Regression

The function k is not used directly, indeed its practical interpretation is unclear. Recall that in the case of Normally distributed data, $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$ a model is specified as $\mu = \mathbf{xfi}$ and the deviation of datum i from this is accounted for by the term ε_i as shown in equation (1).

The non-Normal assumption of generalised linear modelling relates to the distribution of the data—it takes an exponential family distribution. The linearity in parameters is maintained and the model is specified as

$$g\left(b_i(\theta_i)\right) = g(\mu_i) \overset{\text{def}}{=} \mathbf{x}_i\mathbf{fi}$$

where g is called the link function. If $\theta_i = \mathbf{x}_i\mathbf{fi}$ then the link function is, clearly, the inverse of b. However, if for example a non-linear parameterisation is required then using $\theta_i = \ln(\mathbf{x}_i\mathbf{fi})$ could be used in order to achieve the necessary linearity.

In principle both the relationship between θ_i and $\mathbf{x}_i\mathbf{fi}$ and the link function are chosen at will to reflect the practicalities of the model being fitted—an example of where statistical modelling departs from strict mathematical science. In the most simple case $\theta_i = \mathbf{x}_i\mathbf{fi}$ and the link function is the inverse of b; this is called the natural link. An example of this is the Poisson distribution.

$$f(y, \lambda) = \exp\left(y\ln\lambda - \lambda - \ln y!\right)$$

giving $b(\theta) = e^\theta$ and $\theta = \ln\lambda$. Hence $\theta_i = \mathbf{x}_i\mathbf{fi}$ gives $\ln\lambda = \mathbf{x}_i\mathbf{fi}$. The probability density function has now become

$$f(y,\mathbf{fi}) = \exp\left(-e^{\mathbf{x}_i\mathbf{fi}}\right)\left(e^{\mathbf{x}_i\mathbf{fi}}\right)^y \frac{1}{y!}$$

From this the likelihood can be calculated and maximum likelihood estimator found for the parameters **fi**. However, there is no analytic solution to the maximum likelihood equations.

(29.1.5) Goodness Of Fit

Assessing the accuracy of a model for the obtained data is easy in the case of Normally distributed data because analytic expressions are available for the estimated parameters and and can be assessed using the χ^2 and $\mathcal{F}$, distributions.

These results do not hold in general, but the availability of likelihoods suggests the use of a Wald test—likelihood ratio. As the saturated model (with all n parameters) fits the data perfectly it makes sense to compare against this. Let

$$\begin{aligned}\tilde{} &= \begin{pmatrix}\theta_1 & \theta_2 & \dots & \theta_n\end{pmatrix}\\ \hat{\mathbf{fi}} &= \begin{pmatrix}\beta_0 & \beta_1 & \dots & \beta_{p-1}\end{pmatrix}\end{aligned}$$

where in the case of the actual model ` is calculated from **fi**. Hence

$$2\ln\lambda = 2\ln\left(\frac{L\left(\mathbf{y},\tilde{}\right)}{L\left(\mathbf{y},\hat{\mathbf{fi}}\right)}\right) = 2\left(l\left(\mathbf{y},\tilde{}\right) - l\left(\mathbf{y},\hat{\mathbf{fi}}\right)\right)$$

This quantity has an approximate χ^2 on $n-p$ degrees of freedom when the model under test is true. This gives rise to a standard measure of goodness of fit, the defiance.

$$D\left(\mathbf{y},\mathbf{fi}\right) = 2\phi\left(l\left(\mathbf{y},\tilde{}\right) - l\left(\mathbf{y},\hat{\mathbf{fi}}\right)\right)$$

Now, observing data gives rise to a (maximum likelihood) estimate $\hat{\mathbf{fi}}$ for **fi**. From this $\hat{}$ can be calculated and $b'\left(\hat{\theta}_i\right) = \mu_i$. Also, $\tilde{}$ is completely determined, and $b'\left(\tilde{\theta}_i\right) = y_i$. Now, the likelihood function is given by

$$L\left(\mathbf{y},`\right) = \prod_{i=1}^{n} f(y_i,\theta_i) = \prod_{i=1}^{n}\exp\left(\frac{y_i\theta_i - b_i(\theta_i)}{\phi} - c(\phi,y_i)\right)$$

Hence the deviance is given by

$$\begin{aligned}D\left(\mathbf{y},\mathbf{fi}\right) &= 2\phi\left(l\left(\mathbf{y},\tilde{}\right) - l\left(\mathbf{y},\hat{\mathbf{fi}}\right)\right)\\ &= 2\phi\sum_{i=1}^{n}\left(\frac{y_i\tilde{\theta}_i - b_i(\tilde{\theta}_i)}{\phi} - c(\phi,y_i) - \frac{y_i\hat{\theta}_i - b_i(\hat{\theta}_i)}{\phi} + c(\phi,y_i)\right)\\ &= 2\sum_{i=1}^{n}\left(y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - \left(b_i(\tilde{\theta}_i) - b_i(\hat{\theta}_i)\right)\right)\end{aligned}$$

This is the exponential family deviance formula and has an approximate χ^2_{n-p} distribution when the model is true. A common rule of thumb is to reject the model if the deviance lies above the 60% point on its distribution or below the $\frac{1}{2}$% point. This lower limit is set because if the model fits too well it is likely that it has too many parameters.

An alternative measure of goodness of fit is Pearson's χ^2 statistic which may be calculated as

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

where V is the variance function, $V(\theta) = b''(\theta)$. This statistic can be used in another way. The statistic $\frac{X^2}{n-p}$ estimates ϕ, but if this estimate is not near 1 then the Poisson or Binomial model could be wrong.

Deviance On The Normal Distribution

Reassuringly the deviance is consistent with error assessments for Normal linear regression. For the Normal distribution

$$f(y,\mu) = \exp\left(\frac{-y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right)\right)$$

giving

$$\phi = \sigma^2 \qquad \theta = \mu \qquad b(\theta) = \frac{\theta^2}{2}$$

Hence use the identity link $\theta = x_i\beta$ and the deviance is then given by

$$\begin{aligned} D(\mathbf{y},\beta) &= 2\sum_{i=1}^{n}\left(y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - \left(b_i(\tilde{\theta}_i) - b_i(\hat{\theta}_i)\right)\right) \\ &= 2\sum_{i=1}^{n} y_i\left(y_i - \hat{\mu}_i\right) - \left(\frac{y_i^2}{2} - \frac{\mu_i^2}{2})\right) \\ &= \sum_{i=1}^{n} y_i^2 - 2y_i\hat{\mu}_i + \hat{\mu}_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 \end{aligned}$$

This is the usual residual sum of squares.

Analysis Of Deviance

It is usual to build models sequentially (yielding a sequence of models) by adding in one effect at a time. 'One effect' may entail many parameters and every time a new effect is added into the model the deviance but also the degrees of freedom (of the deviance) decreases. A medium must, therefore, be struck between explaining deviance and the use of parameters.

Differences in deviance are themselves χ^2 distributed, and a good rule of thumb is to reject a more complicated model if the difference in deviance is less than 90% significant on the χ^2 distribution for the more complicated model.

(29.1.6) Categorical Factors & Uses In Model Building

Categorical Factors

Thus far measured covariates have been of concern when explaining deviance. However, besides these one may also consider categorical factors such as eye colour where $y_{ijk} = \alpha_i + \beta_j + \varepsilon_{ijk}$ where the subscript k allows more than one datum in each group*.

Let A and B be factors with a and b levels respectively. A possible use of these in a sequence of models is

*Unlike the Normal case there is no requirement to have the same number of data in each group.

1. Fit A alone, denoted A. In this case $g_i = \mu + \alpha_i$.
2. Fit A and B, denoted $A + B$. In this case $g_{ij} = \mu + \alpha_i + \beta_j$.
3. Fit both A and B and also their interaction, denoted $A + B + A.B$. In this case $g_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$.

A fourth model is also available, called the nested model and denoted A/B or $A + A.B$. In the interaction model $A + B + A.B$ the factor B takes the same levels at each level of A. In the nested[†] model B may take different levels at different levels of A.

Modelling A Single Factor In Glim

Consider a single factor A with levels $A_1, A_2, \ldots, A_a$. If there are n_i observations at level A_i then the data may be summarised as shown in Table 29.1.6 where g_{ij} may be replaced by $\mu + \alpha_i$, say.

A_1	A_2	$\ldots$	A_a
g_{11}	g_{21}	$\cdots$	g_{a1}
g_{12}	g_{22}	$\cdots$	g_{a2}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
g_{1n_1}	g_{2n_2}	$\cdots$	g_{an_a}

Table 1: Representation of a single factor.

Such data would be read into Glim with the ys in the first column, say, and indices $1, 2, \ldots, a$ as appropriate in the second column, corresponding to the level of A of each datum. The factor A is identified by means of the command `$factor A a$`. To accommodate this Glim defines the dummy variables $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_a$ such that $u_{ij} = 1$ whenever y_{ij} has level i. The model may therefore be expressed as

$$\mathbf{g} = \mu\mathbf{1} + \alpha_2\mathbf{u}_2 + \cdots + \alpha_a\mathbf{u}_a$$

where $\alpha_1 = 0$ is instead of imposing the constraint $\sum_{i=1}^{a} \alpha_a = 0$. This gives rise to the name "starting point" constraint as opposed to "symmetric" constraint.

A single factor can have useful applications in testing the goodness of fit of a model. Say a number of data are observed at each level of the covariate x. For each x_i introduce a level of a factor A. Fit the model $y_i = \alpha + \beta x_i$ and then add in the group factor using `$fit +A$`. A χ^2 test will determine whether the factor has an appreciable effect; if it does then the linear model may be insufficient. This is called a pure error lack of linear fit test, and in a similar way lack of quadratic fit etc. tests can be performed.

Modelling Cross Classified Factors In Glim

For cross classified factors $g_{ij} = \mu + \alpha_i + \beta_j$ which has the notation $A + B$. In Glim this is represented as

$$\mathbf{g} = \mu\mathbf{1} + \alpha_2\mathbf{u}_2 + \cdots + \alpha_a\mathbf{u}_a + \beta_1\mathbf{v}_1 + \cdots + \beta_b\mathbf{v}_b$$

Using the constraint $\alpha_1 = 0$ and $\beta_1 = 0$—a corner point parameterisation–gives rise to the scheme shown in Table 29.1.6.

For input to Glim the data should be in one column with the two factor levels in two other columns. Alternatively factor levels can be generated in Glim using the 'generate levels' command `$calc f=%gl(p,q)$`

[†] Graphically the interaction model is a simple table. The interaction model may be thought of as the same table but with a block diagonal form.

	A_1	A_2	$\dots$	A_a
B_1	μ	$\mu + \alpha_2$	$\dots$	$\mu + \alpha_a$
B_2	$\mu + \beta_2$	$\mu + \alpha_2 + \beta_2$	$\dots$	$\mu + \alpha_a + \beta_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
B_b	$\mu + \alpha_1 + \beta_b$	$\mu + \alpha_2 + \beta_b$	$\dots$	$\mu + \alpha_a + \beta_b$

Table 2: Scheme for two cross classified factors.

which produces the numbers 1 to p in blocks of q. Hence for the factors A and B the commands `$calc fa=%gl(a,1)$` and `$calc fb=%gl(b,a)$` then `$factor fa a : fb b$` could be used.

It is sometimes the case that a particular factor level or coefficient of a covariate should be 1. This can be achieved using `$offset b$` where b is the coefficient to be set to 1.

Modelling Interacting Factors In Glim

The cross classified model may be expanded upon by supposing that at each level of A and of B another parameter can be added to explain something that happens when the factors interact. The model is therefore

$$g_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

It would be usual to impose the symmetric constraints

$$\sum_{i=1}^{a} \alpha_i = 0 \qquad \sum_{j=1}^{b} \beta_j = 0 \qquad \sum_{i=1}^{a} (\alpha\beta)_{ij} = 0 \qquad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0$$

However, Glim uses the corner point constraint $\alpha_1 = \beta_1 = (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0$. Again Glim introduces dummy variables and this time the model may be expressed as

$$\mathbf{g} = \mu\mathbf{1} + \alpha_2\mathbf{u}_2 + \cdots + \alpha_a\mathbf{u}_a + \beta_1\mathbf{v}_1 + \cdots + \beta_b\mathbf{v}_b + (\alpha\beta)_{22}\mathbf{u}_2\mathbf{v}_2 + \cdots + (\alpha\beta)_{ab}\mathbf{u}_a\mathbf{v}_b$$

where $\mathbf{u}_i\mathbf{u}_j$ in the kth position the product of the kth elements of $\mathbf{u}_i$ and $\mathbf{v}_j$. The effect of this is to 'switch the right $(\alpha\beta)$ on or off'. An interacting cross classification of two factors may be represented as shown in Table 29.1.6.

	A_1	A_2	$\dots$	A_a
B_1	μ	$\mu + \alpha_2$	$\dots$	$\mu + \alpha_a$
B_2	$\mu + \beta_2$	$\mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22}$	$\dots$	$\mu + \alpha_a + \beta_2 + (\alpha\beta)_{a2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
B_b	$\mu + \alpha_1 + \beta_b$	$\mu + \alpha_2 + \beta_b(\alpha\beta)_{2b}$	$\dots$	$\mu + \alpha_a + \beta_b + (\alpha\beta)_{ab}$

Table 3: Scheme for two cross classified factors with interaction.

Such a model can be fitted in Glim using the command `$fit A+B+A.B$` which can of course be used to produce a sequence of models. Alternatively the interaction model can be fitted directly using `$fit A*B$`.

Modelling Nested Factor Models In Glim

A nested factor model is a little like an interacting cross classified model except that some of the parameters are not used and the rest are used differently.

Let A be the first factor and have a levels. At each level of A the factor B takes each of b levels. However, each of the b levels of B is different under each of the a levels of A.

Example 7 *A farmer plants three different varieties of winter barley and plants each variety in two different fields. When in August the crop is harvested it is of interest as to which variety has yielded better: this is factor A and has 3 levels. However, each variety may be further classified as to having grown in one of two fields: this is factor B and takes 2 levels for each level of A but for each level of A the two levels of B are different.*

The notation for such a model is

$$g_{ij} = \mu + \alpha_i + (\alpha\beta)_{i/j}$$

and is interpreted to mean that

$$\begin{array}{rl} \text{when } i = 1 & j = 1, 2, \ldots, b_1 \\ i = 2 & j = b_1 + 1, b_1 + 2, \ldots, b_1 + b_2 \\ \vdots & \vdots \\ i = a & j = 1 + \sum_{i=1}^{a-1} b_i, \ldots, \sum_{i=1}^{a} b_i \end{array}$$

Note there can be a different number of levels of B under each level of A. Glim uses the dummy variables

$$\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_a \quad \text{and} \quad \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_{\sum_{i=1}^{a} b_i}$$

Observe that unlike the interacting cross classified parameterisation the vector $\mathbf{u}_1$ is now used. This gives the representation

$$\begin{aligned} \mathbf{g} = \mu\mathbf{1} &+ (\alpha_2\mathbf{u}_2 + \alpha_3\mathbf{u}_3 + \cdots + \alpha_a\mathbf{u}_a) \\ &+ \alpha_1\mathbf{u}_1 \cdot (\beta_2\mathbf{v}_2 + \cdots + \beta_{b_1}\mathbf{v}_{b_1}) + \ldots \\ &+ \alpha_a\mathbf{u}_a \cdot \left(\beta_{\sum_{i=1}^{a-1} b_i}\mathbf{v}_{2+\sum_{i=1}^{a-2} b_i} + \cdots + \beta_{\sum_{i=1}^{a} b_i}\mathbf{v}_{\sum_{i=1}^{a} b_i}\right) \end{aligned}$$

In Glim the levels of B under A begin numbering at 1, so under the ith level of A the indices of B will be $1, 2, \ldots, b_i$. This manor is reminiscent of a two way cross classification, but is in fact quite distinct. This model is fitted in Glim using the command `$fit A/B$` or equivalently `$fit A+A.B$`.

(29.1.7) Fitting Categorical Factors With Covariates

More generally it will be required to fit both categorical factors and continuous covariates. Producing a regression for each factor will give rise to 'several straight lines' and from there it can be assessed whether the gradients and intercepts are the same or different.

Consider data depending on one covariate and one factor, so it may be summarised as shown in Table 29.1.7.

Issuing the command `$factor A 2$` makes Glim define the dummy covariate $\mathbf{u}_2 = (0, 0, \ldots, 01, 1, \ldots, 1)^T$ to give linear part $\mathbf{g} = \mu\mathbf{1} + \alpha_2\mathbf{u}_2$.

Separate Lines

Perhaps the most general model of this kind is when a separate regression is done for each level of the factor. However, the error variance quoted is for both regressions: 1 regression fits 2 lines.

Y	X	A
y_{11}	x_{11}	1
y_{12}	x_{12}	1
$\vdots$	$\vdots$	$\vdots$
y_{1n_1}	x_{1n_1}	1
y_{21}	x_{21}	2
y_{22}	x_{22}	2
$\vdots$	$\vdots$	$\vdots$
y_{2n_2}	x_{2n_2}	2

Table 4: Data for a factor and a covariate.

This is called the separate lines model and can be fitted with the command `$fit A*X$` which fits the model

$$\mathbf{g} = \mu\mathbf{1} + \alpha_2\mathbf{u}_2 + \beta X + \alpha_2\mathbf{u}_2 \cdot (\beta X)$$

or equivalently

$$\begin{aligned} \mathbf{y} &= \mu\mathbf{1} + \alpha_2\mathbf{u}_2 + \beta X + (\alpha\beta)_2 X\mathbf{u}_2 + '' \\ \text{giving } y_{1i} &= \mu + \beta x_{1i} + \varepsilon_{1i} \\ \text{and } y_{2i} &= \mu + \alpha_2 + (\beta + (\alpha\beta)_2)\, x_{2i} + \varepsilon_{2i} \end{aligned}$$

Thus, as the name suggests, the lines are separate in so much as they have different gradients and intercepts.

No Interaction

The no interaction model is also called the parallel lines model. It is fitted in Glim with the command `$fit A+X$` giving rise to the representation

$$\begin{aligned} \mathbf{y} &= \mu\mathbf{1} + \alpha_2\mathbf{u}_2 + X\beta + '' \\ \text{so } y_{1i} &= \mu + \beta x_{1i} + \varepsilon_{1i} \\ \text{and } y_{2i} &= \mu + \alpha_2 + \beta x_{2i} \end{aligned}$$

Nested Interaction

The nested interaction model has common intercepts. It is fitted using the command `$fit X/A$` giving rise to the parameterisation

$$\begin{aligned} \mathbf{y} &= \mu\mathbf{1} + \beta X + (\beta X) \cdot (\alpha_2\mathbf{u}_2) + '' \\ &= \mu\mathbf{1} + \beta X + (\alpha\beta)_2\mathbf{u}_2 X + '' \\ \text{giving } y_{1i} &= \mu + \beta x_{1i} + \varepsilon_{1i} \\ \text{and } y_{2i} &= \mu + \beta x_{2i} + (\alpha\beta)_2 x_{2i} + \varepsilon_{2i} \end{aligned}$$

It is certainly reassuring that each of the factor parameterisation model have such clear geometrical interpretations.

(29.2) Computational Estimation Theory

(29.2.1) Maximum Likelihood Estimation

Seeking To Maximise The Likelihood

As might be expected, maximum likelihood is the preferred method for parameter estimation in generalised linear models. However, the absence of analytic solutions to the derivative of the likelihood makes the process rather more difficult.

Suppose data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are observed (with covariates $\mathbf{x}_i$ for each y_i) and have associated probability density functions

$$f(y_i, \theta_i) = \exp\left(\frac{y_i\theta_i - b_i(\theta_i)}{\phi} - c(y_i, \phi)\right)$$
$$\text{with } g(\mu_i) = \mathbf{x}_i\mathbf{fi} \text{ where } \mu_i = b_i'(\theta_i)$$

The logarithm of the likelihood is a function of $\mathbf{fi}$ and is given by

$$l(\mathbf{fi}, \mathbf{y}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b_i(\theta_i)}{\phi} - c(y_i, \phi) \tag{8}$$

Since this is a function of the vector $\mathbf{fi}$ vector calculus is now used to 'differentiate'. Suppose $\mathbf{fi}$ is $p \times 1$ then let

$$\mathbf{U}(\mathbf{fi}) = \frac{\partial l}{\partial \mathbf{fi}} = \begin{pmatrix} \frac{\partial l}{\partial \beta_0} \\ \vdots \\ \frac{\partial l}{\partial \mathbf{fi}_{p-1}} \end{pmatrix}$$

For maximum likelihood estimation the equation $\mathbf{U} = \mathbf{0}$ must be solved—this corresponds to solving p equations. Let l_i be the ith term of equation (8), so by the chain rule for partial differentiation

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}$$

These are relatively easy to calculate, and are done so as follows

$$\begin{aligned}
\frac{\partial l_i}{\partial \theta_i} &= \frac{y_i - b_i'(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi} \\
\frac{\partial \mu_i}{\partial \theta_i} &= b_i''(\theta_i) = \frac{\operatorname{var} Y_i}{\phi} \\
\frac{\partial \mu_i}{\partial \beta_j} &= \frac{\partial \mu_i}{\partial g_i}\frac{\partial g_i}{\partial \beta_j} \\
&= \frac{\partial(\mathbf{x}_i\mathbf{fi})}{\partial \beta_j} g^* = x_{ij}g_i^* \quad \text{where } g_i^* = \frac{1}{\frac{\partial g_i}{\partial \mu_i}} \\
\text{hence } \frac{\partial l_i}{\partial \beta_j} &= \frac{y_i - \mu_i}{\phi}\frac{\phi}{\operatorname{var} Y_i} x_{ij}g_i^* \\
&= \frac{(y_i - \mu_i)x_{ij}}{\operatorname{var} Y_i} g_i^*
\end{aligned}$$

This is the ith term of the log likelihood sum for the jth element of $\mathbf{U}$. Hence

$$\begin{aligned}
\frac{\partial l_i}{\partial \mathbf{fi}} &= \frac{(y_i - \mu_i)\mathbf{x}_i^T}{\operatorname{var} Y_i} g_i^* \\
\text{giving } \mathbf{U}(\mathbf{fi}) = \frac{\partial l}{\partial \mathbf{fi}} &= \sum_{i=1}^{n} \frac{(y_i - \mu_i)\mathbf{x}_i^T}{\operatorname{var} Y_i} g_i^* \\
&= \begin{pmatrix} \mathbf{x}_1^T & \mathbf{x}_2^T & \dots & \mathbf{x}_n^T \end{pmatrix} \begin{pmatrix} \frac{(y_1-\mu_i)}{\operatorname{var} Y_1} g_1^* \\ \vdots \\ \frac{(y_n-\mu_i)}{\operatorname{var} Y_n} g_n^* \end{pmatrix} \\
&= X^T W \mathbf{ffi} \qquad (9) \\
\text{where } (X)_{ij} = x_{ij} \qquad (W)_{ii} &= \frac{(g_i^*)^2}{\operatorname{var} Y_i} \qquad \delta_i = (y_i - \mu_i)\frac{1}{g_i^*}
\end{aligned}$$

Note the off diagonal entries of W are zero. $\mathbf{U}(\mathbf{fi})$ is called the score vector for $\mathbf{fi}$ and must now be solved equated to the zero vector. However, the equations are not linear and must be solved numerically or by some iterative process.

The Information Matrix

The information matrix is of use, not least to shorten many expressions. As usual

$$I(\mathbf{fi}) = -\,\mathbb{E}\left(\frac{\partial^2 l}{\partial \mathbf{fi} \partial \mathbf{fi}^T}\right)$$

which can be shown in the usual way. Using equation (9) the expression for the information matrix becomes

$$I(\mathbf{fi}) = \mathbb{E}\left(\frac{\partial l}{\partial \beta} \cdot \frac{\partial l}{\partial \mathbf{fi}^T}\right) = \mathbb{E}\left(\mathbf{U}\mathbf{U}^T\right) = \mathbb{E}\left(X^T W \mathbf{ffi}\mathbf{ffi}^T W X\right)$$

noting that $W = W^T$. The only random variable in this expression is $vtr\delta$, so that $\mathbb{E}\,(\mathbf{ffi}\mathbf{ffi}^T)$ is of interest.

$$\begin{aligned}
\mathbb{E}\left(\mathbf{ffi}\mathbf{ffi}^T\right) &= \begin{cases} \mathbb{E}\left((y_i - \mu_i)^2 (g_i^*)^2\right) & \text{for diagonal terms} \\ \mathbb{E}\left((y_i - \mu_i)(y_j - \mu_j) g_i^* g_j^*\right) & \text{for off diagonal terms} \end{cases} \\
&= \begin{cases} \left(g_i^*\right)^2 \operatorname{var} Y_i & \text{for diagonal terms} \\ 0 & \text{for off diagonal terms, by independence} \end{cases}
\end{aligned}$$

Hence $\mathbb{E}\,(\mathbf{ffi}\mathbf{ffi}^T) = W^{-1}$ so that

$$I(\mathbf{fi}) = X^T W W^{-1} W X = X^T W X$$

The information matrix gives, asymptotically, the variances of the estimate $\hat{\mathbf{fi}}$ of $\mathbf{fi}$ by $\operatorname{var} \hat{\mathbf{fi}} \approx I^{-1}(\hat{\mathbf{fi}})$. These values are given by Glim for the variance estimates of the parameters.

Scoring

At last the equation $\mathbf{U}(\mathbf{fi}) = \mathbf{0}$ is solved. This is done using a vector form of the Newton-Raphson[‡] approximation method.

[‡] Newton-Raphson solves the equation $f(x) = 0$ starting at x_0 by approximating $f(x)$ by the straight line with gradient $f'(x_0)$ that intercepts f at $(x_0, f(x_0))$ and estimating the root of f by where this straight line intercepts the x axis, x_1. The process is then repeated from x_1.

Let $\hat{\mathbf{fi}}_{(1)}$ be the first estimate for $\mathbf{fi}$. Perform a Taylor expansion of $\mathbf{U}(\mathbf{fi})$ gives

$$\frac{\partial l}{\partial \mathbf{fi}} = \frac{\partial l(\hat{\mathbf{fi}}_{(1)})}{\partial \mathbf{fi}} + \left(\mathbf{fi} - \hat{\mathbf{fi}}_{(1)}\right) \frac{\partial^2 l(\hat{\mathbf{fi}}_{(1)})}{\partial \mathbf{fi} \partial \mathbf{fi}^T} + \ldots$$

Using only these first two terms,

$$\begin{aligned} \mathbf{fi} - \hat{\mathbf{fi}}_{(1)} &= \left(-\frac{\partial^2 l(\hat{\mathbf{fi}}_{(1)})}{\partial \mathbf{fi} \partial \mathbf{fi}^T}\right) \frac{\partial l(\hat{\mathbf{fi}}_{(1)})}{\partial \mathbf{fi}} \\ \text{so } \hat{\mathbf{fi}}_{(m)} &= \hat{\mathbf{fi}}_{(m-1)} \left(-\frac{\partial^2 l(\hat{\mathbf{fi}}_{(m-1)})}{\partial \mathbf{fi} \partial \mathbf{fi}^T}\right) \frac{\partial l(\hat{\mathbf{fi}}_{(m-1)})}{\partial \mathbf{fi}} \end{aligned}$$

The method of scoring, or statistical modification, is to replace the second order partial derivative in the iteration equation with the expected information matrix evaluated at $\hat{\mathbf{fi}}_{(m-1)}$. Hence

$$\begin{aligned} \hat{\mathbf{fi}}_{(m)} &= \hat{\mathbf{fi}}_{(m-1)} + I^{-1}\left(\hat{\mathbf{fi}}_{(m-1)}\right) \frac{\partial l(\hat{\mathbf{fi}}_{(m-1)})}{\partial \mathbf{fi}} \\ I\left(\hat{\mathbf{fi}}_{(m-1)}\right) \hat{\mathbf{fi}}_{(m)} &= I\left(\hat{\mathbf{fi}}_{(m-1)}\right) \hat{\mathbf{fi}}_{(m-1)} + \mathbf{U}\left(\hat{\mathbf{fi}}_{(m-1)}\right) \\ \left(X^T W_{(m-1)} X\right) \hat{\mathbf{fi}}_{(m)} &= \left(X^T W_{(m-1)} X\right) \hat{\mathbf{fi}}_{(m-1)} + X^T W_{(m-1)} \mathbf{ffi} \\ &= X^T W_{(m-1)} Z_{(m-1)} \qquad (10) \\ \text{where } Z_{(m-1)} &= X \hat{\mathbf{fi}}_{(m-1)} + \mathbf{ffi}_{(m-1)} \end{aligned}$$

Quite why this equation is useful will become apparent in the next section.

(29.2.2) Weighted Least Squares Estimation

Weighted Least Squares

Weighted least squares is much the same as normal least squares, but with the obvious expectation that the terms in the sum are weighted, hence

$$l(\mathbf{fi}) = \sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i \mathbf{fi})^2$$

is the quantity to be minimised. Using matrix and vector notation this may be written as $l(\mathbf{fi}) = (\mathbf{y} - X\mathbf{fi})^T W (\mathbf{y} - X\mathbf{fi})$. Now using the vector differentiation results

$$\frac{\partial \mathbf{fi}^T}{\partial \mathbf{fi}} = I \quad \text{and} \quad \frac{\partial}{\partial \mathbf{fi}}\left(\mathbf{a}^T W \mathbf{a}\right) = 2 \frac{\partial \mathbf{a}^T}{\partial \mathbf{fi}} W \mathbf{a}$$

on the vector representation of the sum of squares gives

$$\frac{\partial l(\mathbf{fi})}{\partial \mathbf{fi}} = 2X^T W (\mathbf{y} - X\mathbf{fi})$$

so that at the minimum

$$\begin{aligned} 2X^T W(\mathbf{y} - X\mathbf{fi}) &= 0 \\ \text{or equivalently } 2X^T W \mathbf{y} &= 2X^T W X \mathbf{fi} \\ \text{or equivalently } X^T W \mathbf{ffi}(\mathbf{fi}) &= 0 \qquad (11) \\ \text{where } \mathbf{ffi}(\mathbf{fi}) &= \mathbf{y} - X\mathbf{fi} \end{aligned}$$

Equation (11) has the same form as equation (9) except the choice of **ffi** admits the analytic solution

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

Observe that putting $W = I_n$ gives the usual maximum likelihood estimate, as is shown in Chapter **??** on general linear modelling. Similarly the proper generalisation of the estimates for the variance-covariance matrix of the parameters is $\sigma^2(X^T W X)^{-1}$ and σ^2 is estimated as

$$\frac{1}{n-p}\left((\mathbf{y} - X\mathbf{ffi})^T W(\mathbf{y} - X\mathbf{ffi})\right)$$

Using least squares no mention of the exponential family of distributions has been made. The least squares method may therefore be used when the exponential family assumption is not true. However, equations (11) and (9) having the same form means that least squares will give the maximum likelihood estimate under the exponential family assumption. This is important when modelling strays from the exponential distribution, as is discussed in Section 29.2.3.

Interpreting Weighted Least Squares Estimation

Return now to consider the maximum likelihood estimation process and in particular equation (11) where

$$Z_{(m-1)} = X\hat{\mathbf{ffi}}_{(m-1)} + \mathbf{ffi}_{(m-1)} \quad \text{and} \quad \delta_i = (y_i - \mu_i)\frac{\partial g_i}{\partial \mu_i}$$

Consider now a Taylor expansion of the link function about μ so

$$g(y) \approx g(\mu) + (y - \mu)\frac{\partial g}{\partial \mu}$$

from which the variance may be calculated as

$$\operatorname{var}(g(Y)) = \left(\frac{\partial g}{\partial \mu}\right)^2 \operatorname{var} Y$$

This is the inverse of the weighting elements of W, and it should be noted that this is not constant.

(29.2.3) Dispersion

(29.3) Model Building And Assessment

Model Building

Model building is more of an art than a science, many of the decisions being down to experience and judgement. Which explanatory variables or factors to use and which are most important is the first decision to make, and as to how they are to be fitted: which will interact etc. Some rules of thumb have already been mentioned.

1. A reasonable model should have deviance greater than $\frac{1}{2}$% and 60% on the corresponding χ^2 distribution.
2. The current model is likely to be under fitted (not enough parameters) if the deviance is above the 90% point on the corresponding χ^2 distribution.

3. If adding a parameter causes a decrease in deviance of less than 10% on the corresponding χ^2 distribution then that parameter may not be necessary.

4. If removing a parameter causes in increase in deviance of more than 90% on the corresponding χ^2 distribution then that parameter is likely to have an important effect.

However, if different combinations of covariates are used then the change in deviance when effects are added and removed will change. This is why it is important to judge the relative importance of the covariates and fit the most important ones first.

There is also some choice in what link function to use. It is usual to use $\theta_i = \mathbf{x}_i\mathbf{fi}$ and deduce the link function, but if it is preferable to multiply parameters together (rather than add) then a logarithm will have to be taken giving $\ln \theta_i = \mathbf{x}_i\mathbf{fi}$.

At all times it should be remembered that as few parameters as possible should be used. As fitting categorical factors is very heavy on the use of parameters it may be feasible to combine some of the levels or to convert the category to a continuous covariate e.g. age may be represented by a covariate rather than categories.

Goodness Of Fit Using Pearson's χ^2 Statistic

An alternative to deviance when assessing the goodness of fit of a model is Pearson's generalised X^2 statistic

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

where $V(\hat{\mu}_i)$ is the estimate of the variance function of the exponential family distribution under consideration, i.e. the estimate of $\operatorname{var} Y_i = \phi b_i''(\theta_i)$. This statistic has an approximate χ^2 distribution with the same degrees of freedom as would have the deviance. The scale parameter ϕ can be estimated by $\frac{X^2}{df}$ rather than using the deviance.

Assessing Residuals

For generalised linear modelling there is no particularly clear definition to take for residuals. The Pearson residuals are used, being calculated as

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i}}$$

These are stored in Glim after fitting a model in the variable `%rs`. These can be displayed using `$display r$` and as a rough check a good model should have the residuals in the range $(-2, 2)$, working from a Normal assumption.

(29.3.1) Problems Using Deviance For Binomial Regression

For the Binomial distribution

$$f(x, p) = \binom{n}{p} p^x (1-p)^{n-x} = \exp\left(x \ln\left(\frac{p}{1-p}\right) + n \ln(1-p) + \ln\binom{n}{p}\right)$$

so that

$$\begin{aligned}
\theta* &= \ln\left(\frac{p}{1-p}\right)\\
(1-p)e^{\theta} &= e^{\theta}\\
e^{\theta} &= p(1+e^{\theta})\\
p &= \frac{e^{\theta}}{1+e^{\theta}}
\end{aligned}$$

Hence an expression for $b(\theta)$ can be found,

$$b(\theta) = -n\ln(1-p) = n(\theta - \ln p) = n\theta - n\theta + n\ln(1+e^{\theta}) = n\ln(1+e^{\theta})$$

The canonical link is chosen which gives $\theta_i = \mathbf{x}_i\mathbf{fi}$. The actual link function is $g(\mu_i)$ such that $g(\mu_i) = \mathbf{x}_i\mathbf{fi}$. Now, $\mu_i = b'(\theta_i)$.

$$\begin{aligned}
b(\theta_i) &= n\ln(1+e^{\theta_i})\\
\text{so } b'(\theta_i) &= \frac{ne^{\theta_i}}{1+e^{\theta_i}}\\
&= np
\end{aligned}$$

Hence since $\mathbf{x}_i\mathbf{fi} = \theta_i = \ln\left(\frac{p_i}{1-p_i}\right)$ this gives

$$g(\mu_i) = \ln\left(\frac{\frac{\mu_i}{n}}{1-\frac{\mu_i}{n}}\right)$$

Consider the k parameter model where **fi** and **x** are used to generate the n θ parameters. In the k parameter case say $\hat{\ } = \begin{pmatrix}\hat{\theta}_1 & \hat{\theta}_2 & \ldots & \hat{\theta}_n\end{pmatrix}$ while the saturated model where all n parameters are used has the parameters $\tilde{\ } = \begin{pmatrix}\tilde{\theta}_1 & \tilde{\theta}_2 & \ldots & \tilde{\theta}_n\end{pmatrix}$. The deviance would then be calculated by the equation

$$\begin{aligned}
D &= 2\sum_{i=1}^{n} y_i(\hat{\theta}_i - \tilde{\theta}_i) + b(\hat{\theta}_i) - b(\tilde{\theta}_i)\\
&= -2\sum_{i=1}^{n}
\end{aligned}$$

(29.4) Contingency Table Frequency Data

(29.4.1) Cross Classified Categorical Response Variables

Description

Thus far the response variables have always been continuous quantities. This is frequently not the case, where the response is categorical, and may depend on categorical covariates. An example is shown in Table 29.4.1, note that any of the three categories could be treat as the response and in fact more than 1 category could be treat as the response.

Modelling Two Categories

First of all consider only 2 categories. Here, as in all cases, the multinomial distribution will be used to assign probabilities of an observation being in a particular cell of the frequency table. Let p_{ij} be the probability

Adversity of school conditions	Low		Medium		High	
Home conditions	Bad	Good	Bad	Good	Bad	Good
Normal	16	7	15	34	5	3
Deviant	1	1	3	8	1	3

Table 5: Classroom behavior data of school children.

that an observation falls into the ith level of category A and the jth level of category B. Suppose that n observations are made then the number of observations in each cell, n_{ij} is of interest. This is modelled by the random variable N_{ij} which has joint distribution function

$$\Pr\left\{N_{ij} = n_{ij}\ \forall i\ \forall j \mid n\right\} = \frac{n!}{\prod_{i=1}^{a}\prod_{j=1}^{b} n_{ij}!}\prod_{i=1}^{a}\prod_{j=1}^{b} p_{ij}^{n_{ij}} \qquad \text{with } \sum_{i=1}^{a}\sum_{j=1}^{b} p_{ij} = 1 \tag{12}$$

For the purposes of modelling structure must be given to the p_{ij} to reduce the number of parameters from ab to something more manageable. Let

$$p_{ij} = \frac{\mu_{ij}}{\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}}$$

then certainly they sum to 1. The means are now parameterised in a log linear way, similar to the Poisson regression case. This is explained in Section 29.4.2. So

$$\ln \mu_{ij} = \phi + \mathbf{x}_{ij}\mathbf{fi}$$

A particular model of much importance uses

$$\ln \mu_{ij} = \phi + \alpha_i + \beta_j$$

hence

$$\begin{aligned} p_{ij} &= \frac{e^{\alpha_i+\beta_j}}{\sum_{i=1}^{a}\sum_{j=1}^{b} e^{\alpha_i+\beta_j}} \\ &= \frac{e^{\alpha_i}}{\sum_{i=1}^{a} e^{\alpha_i}}\frac{e^{\beta_j}}{\sum_{j=1}^{b} e^{\beta_j}} \end{aligned}$$

So $p_{ij} = p_i p_j$ showing that in this parameterisation A and B are independent. Returning to generality

$$\begin{aligned} p_{ij} &= \frac{e^{\phi+\mathbf{x}_{ij}\mathbf{fi}}}{\sum_{i=1}^{a}\sum_{j=1}^{b} e^{\phi+\mathbf{x}_{ij}\mathbf{fi}}} \\ &= \frac{e^{\mathbf{x}_{ij}\mathbf{fi}}}{\sum_{i=1}^{a}\sum_{j=1}^{b} e^{\mathbf{x}_{ij}\mathbf{fi}}} \end{aligned}$$

Conveniently the ϕ has cancelled out. To estimate **fi** maximum likelihood is used.

$$\begin{aligned}
L = \Pr\left\{N_{ij} = n_{ij}\ \forall i\ \forall j \mid n\right\} &= \frac{n!}{\prod_{i=1}^{a}\prod_{j=1}^{b} n_{ij}!}\prod_{i=1}^{a}\prod_{j=1}^{b} p_{ij}^{n_{ij}} \\
l &= \ln n! - \sum_{i=1}^{a}\sum_{j=1}^{b}\ln n_{ij} + \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\left(\mathbf{x}_{ij}\mathbf{fi} - \ln\left(\sum_{i=1}^{a}\sum_{j=1}^{b} e^{\mathbf{x}_{ij}\mathbf{fi}}\right)\right) \\
&= \ln n! - \sum_{i=1}^{a}\sum_{j=1}^{b}\ln n_{ij} + \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\mathbf{x}_{ij}\mathbf{fi} - n\ln\left(\sum_{i=1}^{a}\sum_{j=1}^{b} e^{\mathbf{x}_{ij}\mathbf{fi}}\right) \qquad (13)
\end{aligned}$$

This needs to be minimised, which must be done numerically. Conveniently this too is independent of ϕ.

As already seen the form of $\mathbf{x}_{ij}\mathbf{fi}$ determines the joint distribution of the category variables.

(29.4.2) Tricking Glim

Glim cannot perform the required calculations for such frequency table data. However, it can be tricked into doing it. Discarding the multinomial distribution, assume that each n_{ij} has an independent Poisson distribution with mean μ_{ij}. The likelihood (corresponding to equation 12) is then

$$\begin{aligned}
L(\bar{\ }) &= \prod_{i=1}^{a}\prod_{j=1}^{b}\frac{e^{-\mu_{ij}}\mu_{ij}^{n_{ij}}}{n_{ij}!} \\
l(\bar{\ }) &= \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\ln\mu_{ij} - \mu_{ij} + \ln n_{ij}!
\end{aligned}$$

Now put $\ln\mu_{ij} = \phi + \mathbf{x}_{ij}\mathbf{fi}$ and continue

$$= \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\phi + \mathbf{x}_{ij}\mathbf{fi} - e^{\phi+\mathbf{x}_{ij}\mathbf{fi}} + \ln n_{ij}!$$

Now, ϕ is not a particularly appropriate parameter. Reparameterise therefore using $\tau = \sum_{i=1}^{a}\sum_{j=1}^{b} e^{\phi+\mathbf{x}_{ij}\mathbf{fi}}$ so that $\phi = \ln\tau - \ln\left(\sum_{i=1}^{a}\sum_{j=1}^{b} e^{\mathbf{x}_{ij}\mathbf{fi}}\right)$

$$\begin{aligned}
&= \left(\ln\tau - \ln\left(\sum_{i=1}^{a}\sum_{j=1}^{b} e^{\mathbf{x}_{ij}\mathbf{fi}}\right)\right)\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij} - \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\mathbf{x}_{ij}\mathbf{fi} - \tau \\
&= n\ln\tau - \tau - n\ln\left(\sum_{i=1}^{a}\sum_{j=1}^{b} e^{\mathbf{x}_{ij}\mathbf{fi}}\right) + \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\mathbf{x}_{ij}\mathbf{fi}
\end{aligned}$$

However, the expression in **fi** here is the same as in the multinomial log likelihood, equation (13). Moreover, when differentiated with respect to **fi** both equations will be identical, giving the same estimate for **fi**. The estimate of τ is easy to calculate as it is independent of **fi**, $\hat{\tau} = n$. However, this is of little interest as the analysis given is conditional on n.

From the Poisson point of view the situation now arisen is to model random variables N_{ij} each with a mean

μ_{ij} but conditional on their sum being n. Since the N_{ij} random variables are independent this gives

$$\begin{aligned}
\Pr\left\{ N_{ij} = n_{ij}\ \forall i\ \forall j \mid \sum_{i=1}^{a}\sum_{j=1}^{b} N_{ij} = n \right\} &= \frac{\prod_{i=1}^{a}\prod_{j=1}^{a} \frac{e^{-\mu_{ij}}\mu_{ij}^{n_{ij}}}{n_{ij}!}}{\frac{1}{n!}\exp\left(-\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}\right)\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}\right)^{n}} \\
&= \frac{n!}{\prod_{i=1}^{a}\prod_{j=1}^{a} n_{ij}!}\prod_{i=1}^{a}\prod_{j=1}^{a}\left(\frac{\mu_{ij}}{\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}}\right)^{n_{ij}}
\end{aligned}$$

which is the same as equation (12) with the chosen parameterisation i.e. the parameterisation was chosen so that this Poisson methodology could be applied hence allowing analysis in Glim.

Glim can be tricked into maximising the multinomial likelihood by setting `$error p$` and using `$yvar n_ij$`. The factors can then be fitted in the usual way.

(29.4.3) Parametric Structures For Two Dimensional Tables

As already discussed $\ln \mu_{ij} = \phi + \alpha_i + \beta_j$ gives a model in which the A and B categories are independent which is deduced by showing the joint distribution to be the product of the marginal distributions. For the record the marginal distributions may be found from p_{ij} as follows

$$\begin{aligned}
p_i &= \sum_{j=1}^{b} p_{ij} \\
&= \sum_{j=1}^{b}\frac{\mu_{ij}}{\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}} \\
&= \frac{e^{\alpha_i}\sum_{j=1}^{b} e^{\beta_j}}{\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}} \\
&= \frac{e^{\alpha_i}}{\sum_{i=1}^{a} e^{\alpha_i}} \\
\text{similarly } p_j &= \frac{e^{\beta_j}}{\sum_{j=1}^{b} e^{\beta_j}}
\end{aligned}$$

(29.4.4) Parametric Structures For Three Dimensional Tables

(29.4.5) Full Independence

If the effects A, B, and C are all independent, written $A \perp B \perp C$ then

$$p_{ijk} = \Pr\{A = i\} \times \Pr\{B = j\} \times \Pr\{C = k\}$$

giving rise to the parameterisation

$$\ln \mu_{ijk} = \phi + \alpha_i + \beta_j + \gamma_k$$

Joint Independence

In the joint independence model the joint distribution of A and B is independent of the distribution of C. This is written $(A, B) \perp C$. Hence

$$\begin{aligned} \Pr\{A = i, B = j \mid C = k\} &= \Pr\{A = i, B = j\} \\ \text{so } \Pr\{A = i, B = j, C = k\} &= \Pr\{A = i, B = j\} \times \Pr\{C = k\} \\ \text{suggesting } \ln \mu_{ijk} &= \phi + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} \end{aligned}$$

In this model there are two marginal distributions, one is the distribution of C and the other is the joint distribution of (A, B).

$$\begin{aligned} p_{ij} &= \sum_{k=1}^{c} \frac{\exp\left(\phi + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}\right)}{\sum_{i=1}^{a} \sum_{j=1}^{b} \exp\left(\phi + \alpha_i + \beta_j + (\alpha\beta)_{ij}\right) \sum_{k=1}^{c} e^{\gamma_k}} \\ &= \frac{e^{\phi+\alpha_i+\beta_j+(\alpha\beta)_{ij}}}{\sum_{i=1}^{a} \sum_{j=1}^{b} e^{\phi+\alpha_i+\beta_j+(\alpha\beta)_{ij}}} \end{aligned}$$

Hence the marginal distribution for (A, B) has $\ln \mu_{ij} = \phi + \alpha_i + \beta_j + (\alpha\beta)_{ij}$. Summing over i and j to obtain the marginal distribution of C gives

$$p_k = \frac{e^{\gamma_k}}{\sum_{k=1}^{c} e^{\gamma_k}}$$

Now the distribution of (A, B) conditional on C can be found

$$\begin{aligned} p_{ij|K} &= \frac{p_{ijk}}{p_k} \\ &= \frac{e^{\phi+\alpha_i+\beta_j+(\alpha\beta)_{ij}}}{\sum_{i=1}^{a} \sum_{j=1}^{b} e^{\phi+\alpha_i+\beta_j+(\alpha\beta)_{ij}}} \end{aligned}$$

So the same model as for the marginal distribution of (A, B) is used. This gives $p_{ij|k} = p_{ij}$ showing that (A, B) is independent of C (which was known) so this model is correct.

The independence with C means that the (A, C) and (B, C) marginal distributions will both indicate some kind of independence relationship. However, in the case of (A, C), $p_{ik} \neq p_{ik|j}$

Conditional Independence

In this model A and B are independent but are conditioned on C. This is written $(A \perp B)|C$ or equivalently $(A|C) \perp (B|C)$ giving

$$\begin{aligned} \Pr\{A = i, B = j \mid C = k\} &= \Pr\{A = i \mid C = k\} \times \Pr\{B = j \mid C = k\} \\ \text{so } \Pr\{A = i, B = j, C = k\} &= \Pr\{A = i \mid C = k\} \times \Pr\{B = j \mid C = k\} \times \Pr\{C = k\} \\ &= \frac{\Pr\{A = i, C = k\} \times \Pr\{B = j, C = k\}}{\Pr\{C = k\}} \\ \text{suggesting } \ln \mu_{ijk} &= \phi + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \end{aligned}$$