Notes on Transformations and Generalized Linear Models

W N Venables and Clarice G B Demétrio

2007-08-19

Contents

1 Introduction

These notes are intended to provide an introduction to generalized linear modelling, emphasising the way relationship between the modern theory and the older theory of transformations, out of which the idea developed.

We consider transformations in statistics, however, to be of much more than historical interest. The brief treatment we give here is intended to be as much for their use in contemporary data analysis as for showing the origins of the idea of a generalized linear model.

2 Transformations

2.1 Approximate means and variances

Let Y be a random variable with first two moments

$$\mathrm{E}[Y] = \mu \qquad \text{and} \qquad \mathrm{var}[Y] = \mathrm{E}\left[(Y-\mu)^2\right] = \sigma^2.$$

Now let $U = g(Y)$ be another random variable defined as a function of Y and we need an approximate expression for its first two moments as well. If we can assume that $g(.)$ is smooth and only slowly varying, at least in the region where its argument, Y, is stochastically located, the simplest approach to this problem is to assume that a linear approximation to $g(.)$ near the mean of Y is adequate. Expanding $g(.)$ in a Taylor series gives

$$U = g(Y) = g(\mu) + g'(\mu)(Y-\mu) + \text{“smaller order terms”}$$

Neglecting the smaller order terms gives the approximate expressions

$$\mathrm{E}[U] \approx g(\mu) + g'(\mu)\mathrm{E}\left[(Y-\mu)\right] = g(\mu) \tag{1}$$

$$\mathrm{var}[U] \approx \mathrm{E}\left[\left(U - g(\mu)\right)^2\right] \approx g'(\mu)^2\mathrm{E}\left[(Y-\mu)^2\right] = g'(\mu)^2\sigma^2 \tag{2}$$

Approximate formulae 1 and 2, and extensions to them, are often referred to in statistics as “the delta method”. They are useful in their own right, but they also give some elementary guidance about the possible choices of transformation to achieve various aims.

2.2 Variance stabilising transformations

If the variance of Y is not constant but changes with the mean, that is $\mathrm{var}[Y] = \sigma^2(\mu)$, this can often cause difficulties with both interpretation and analysis. In these cases one possible way around the difficulties might be to transform the response, Y, to a new scale in which the variance is at least approximately constant.

Suppose, then, that we transform the response to $U = g(Y)$. The delta method suggests that if we want the variance of U to be approximately constant, then we should choose $g(.)$ such that

$$\mathrm{var}[g(Y)] \approx g'(\mu)^2\sigma^2(\mu) = k^2$$

where k is a constant. In other words, we should choose $g(.)$ to be any solution of

$$g'(t) = \frac{dg}{dt} = \frac{k}{\sigma(t)}$$

up to changes in location and scale. A convenient solution, then, is

$$g(y) = \int^y \frac{dt}{\sigma(t)}$$

Example 2.1 If Y has a Poisson distribution, $Y \sim \mathrm{Po}(\mu)$, then

$$\mathrm{E}[Y] = \mathrm{var}[Y] = \mu = \sigma^2(\mu)$$

To transform the distribution to approximately constant variance, then, the suggested transform is

$$g(y) = \int^y \frac{dt}{\sigma(t)} = \int^y \frac{dt}{\sqrt{t}} = 2\sqrt{y}$$

Taking the square root was a standard technique in the analysis of count data and towards the middle of the last century much work was done to refine it.

Example 2.2 Suppose S is a Binomial random variable, $S \sim \mathrm{B}(n, \varpi)$, and put $Y = S/n$, the 'proportion of successes'. Then

$$\mathrm{E}[Y] = \varpi = \mu, \qquad \mathrm{var}[Y] = \sigma^2(\mu) = \frac{\mu(1-\mu)}{n}$$

Hence, up to location and scale, the suggested transformation that will approximately stabilise the variance is

$$g(y) = \int^y \frac{dt}{\sigma(t)} = \sqrt{n} \int^y \frac{dt}{\sqrt{t(1-t)}} = \sqrt{n} \sin^{-1} \sqrt{y}$$

Transforming with an 'arc-sine square-root' was a standard technique in the analysis of proportion data and, as in the Poisson case, much work was done to refine it prior to the general adoption of generalised linear modelling alternatives.

Example 2.3 A distribution for which the ratio $\mathrm{cv} = \sigma/\mu = k$ is constant with respect to the mean is said to have "constant coefficient of variation". Since $\sigma^2(\mu) = k^2\mu^2$, the suggested transformation to stabilise the variance is

$$g(y) = \int^y \frac{dt}{\sigma(t)} = \frac{1}{k} \int^y \frac{dt}{t} = \frac{1}{k} \log(y)$$

Hence for such distributions the $\log$ transformation is suggested to make the variance at least approximately constant with respect to the mean.

As an exercise, show that both the gamma and lognormal distributions have constant coefficient of variation, and examine to what extent the log transformation stabilises the variance with respect to the mean.

The gamma distribution has probability density function

$$f_Y(y; \alpha, \phi) = \frac{e^{-y/\alpha} y^{\phi-1}}{\alpha^\phi \Gamma(\phi)}, \qquad 0 < y < \infty$$

The lognormal distribution is defined by transformation. We say Y has a lognormal distribution if $\log Y \sim \mathrm{N}(\mu, \sigma^2)$.

2.3 The Box-Cox family of transformations

Transforming a response to stabilise the variance will, of course, also affect the relationship between the mean and the candidate predictors. In a pioneering paper [Box & Cox(1964)] Box and Cox suggested a method for choosing a transformation that allowed the effect on both the mean and the variance to be taken into account. They considered a family of transformations defined by

$$g(y;\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases} \qquad \text{with} \qquad \frac{dg(y;\lambda)}{dy} = y^{\lambda-1}$$

Note that this includes both the square-root and log transformations, along with other power transformations which are often used in practice, (including the trivial identity transformation).

Now suppose we have a sample of responses and that after transformation it conforms to a linear model specification as follows (with an obvious notation):

$$g(\boldsymbol{y};\lambda) \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$$

The likelihood function for the sample is the distribution of $\mathbf{y}$, namely

$$\log \mathscr{L}(\boldsymbol{\beta}, \sigma^2, \lambda; \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\|g(\mathbf{y};\lambda) - \mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma^2} + \log\prod_{i=1}^n y_i^{\lambda-1}$$

where the final term on the right is the Jacobian factor for the inverse transformation. (This is only an approximate result in general as for most transformations in the family the range is not $-\infty < y < \infty$, but we ignore this here.)

Maximising this with respect to $\boldsymbol{\beta}$ and σ^2 gives the profile likelihood for λ, which by standard results is easily shown to be

$$\log \mathscr{L}^\star(\lambda;\mathbf{y}) = \max_{\boldsymbol{\beta},\sigma^2|\lambda} \log \mathscr{L} = -\frac{n}{2}\log(2\pi/n) - \frac{n}{2}\log\left\{g(\mathbf{y};\lambda)^\mathsf{T}(\mathbf{I} - \mathbf{P}_\mathbf{X})g(\mathbf{y};\lambda)\right\} - \frac{n}{2} + \log\prod_{i=1}^n y_i^{\lambda-1}$$

where $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^-\mathbf{X}^\mathsf{T}$ is the orthogonal projector matrix on to the range of $\mathbf{X}$, and the quantity in braces, $\{\ldots\}$, is the residual sum of squares after regressing the transformed response on $\mathbf{X}$.

As pointed out by Box and Cox, the Jacobian factor can be combined with RSS term in a neat way. Note that

$$\log\prod_{i=1}^n y_i^{\lambda-1} = \frac{n}{2}\log \dot{y}^{2(\lambda-1)}$$

where $\dot{y} = \left(\prod_{i=1}^n y_i\right)^{1/n}$ is the geometric mean of the observations. Now define a slightly modified response as

$$\mathbf{z}(\lambda) = g(\mathbf{y};\lambda) \,/\, \dot{y}^{\lambda-1}$$

Then the profile likelihood for λ may be written

$$\log \mathscr{L}^\star(\lambda;\mathbf{y}) = \text{const.} - \frac{n}{2}\log\left\{\mathbf{z}(\lambda)^\mathsf{T}(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{z}(\lambda)\right\}$$

where "const." does not depend on λ. Hence the profile likelihood may be computed from the residual sum of squares after regressing the constructed response, $\mathbf{z}(\lambda)$ on $\mathbf{X}$.

The strategy suggested by Box and Cox was to use the profile likelihood *not* to estimate λ, but to serve as a guide in the choice of the transformation, using as much contextual information as possible. They suggest fitting a model with the desired mean structure and constant variance initially to the untransformed response variable, and consider the profile likelihood for a series of values of λ. If it is possible for a member of the transformation family to achieve the desired mean structure and constant variance, it will come from the set with high profile likelihood, but the onus is still on the modeller to show that the result has been satisfactorily achieved, using, for example, diagnostic checks.

The standard plot of the profile likelihood, with the formal MLE and likelihood-ratio 95% confidence interval shown can be useful for this purpose. Such a plot is available using the `boxcox` from the `MASS` package.

Example 2.4 The `Cars93` data set from the `MASS` library contains information on the models of cars released in the USA in 1993. Two of the variables given are `MPG.city`, the fuel efficiency in city driving, and `Weight`. Suppose we wish to choose a scale for the fuel efficiency variable so that it can be predicted by a linear regression on the weight of the vehicle.

```
> library(MASS)
> par(mfrow = c(2, 2), cex = 0.7)
> with(Cars93, plot(Weight, MPG.city, col = "green4"))
> FEM.orig <- lm(MPG.city ~ Weight, Cars93)
> plot(fitted(FEM.orig), resid(FEM.orig), col = "navy")
> abline(h = 0, lty = "dashed", col = "red")
> boxcox(FEM.orig, lambda = seq(-1.75, -0.5, len = 10))
```

Figure 1.

The reciprocal transformation, $\lambda = -1$ stands out as both natural in the context, since fuel economy may equally be measured as 'gallons per mile' as by 'miles per gallon', and acceptable from the point of view of the profile likelihood. For scale reasons we choose a minor variation on this, namely `1000/MPG.ciy`, or 'gallons per 1000 miles'.

```
> with(Cars93, plot(Weight, 1000/MPG.city, col = "red"))
```

The results are shown in the final panel of Figure 1. The regression line is visibly straight and the spread about the line much more homogeneous.

Example 2.5 An alternative family of transformations that can be studied in the same way as the Box-Cox family is the "displaced log" family, defined as

$$h(y;\alpha) = \log(y + \alpha)$$

As an exercise, derive the profile likelihood for α given that the transformed response has a linear model of the same form as that described for the Box-Cox family.

The `MASS` library function `logtrans` may be used in a similar way to the `boxcox` function. An example from the Quine follows, with the output shown in Figure 2.

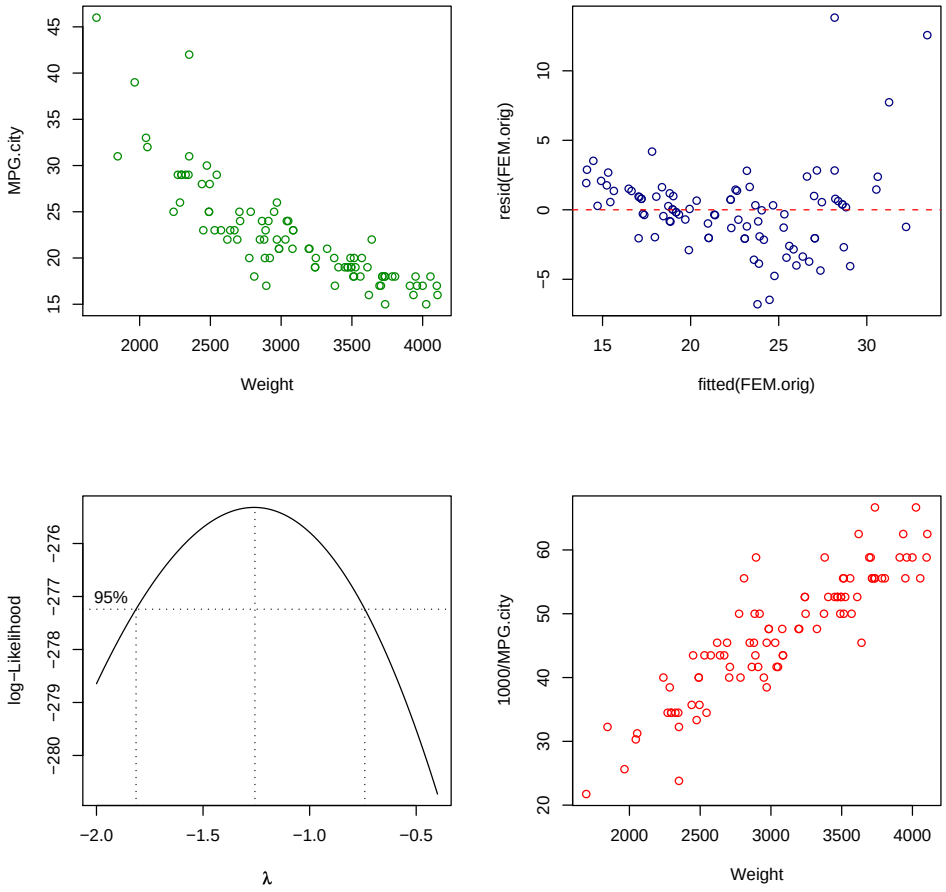

Figure 1: Data and diagnostic plots for the fuel efficiency model.

```
> library(MASS)
> logtrans(Days ~ Eth * Lrn * Age * Sex, data = quine)
```

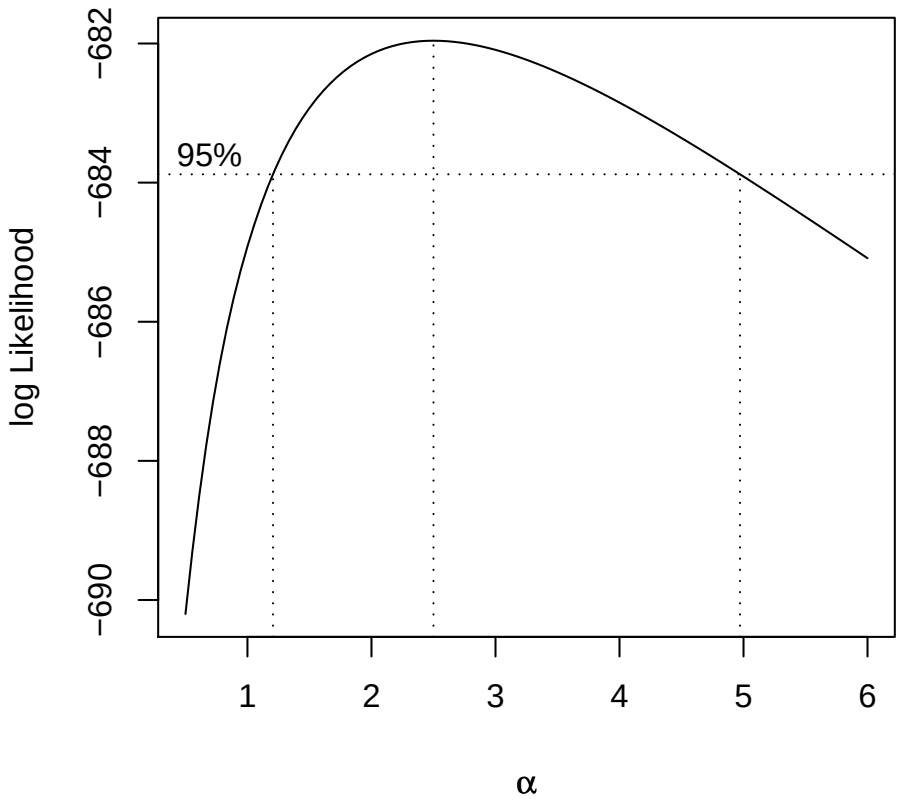

Figure 2: Profile likelihood for the α parameter in a displaced log transform with the Quine data.

3 Introduction to generalized linear models

The idea of a generalized linear model (GLM) developed from the practice in statistics of transforming the response variable to provide a convenient scale for analysis. Transforming the scale of the response was a simple device, but it had two main drawbacks, namely

- Transforming the response variable could be used to simplify the mean structure or to stabilise the variance, but often not both.
- If an analysis were done in a transformed scale, it was often difficult to make predictions, for example, in the original scale, which was often necessary.

The Box-Cox proposals were aimed at achieving a reasonable compromise between variance stability and simplicity of mean structure with a single transformation, but it would clearly be preferable to have separate mechanisms for simplifying mean structure and stabilising variance.

GLMs get around these problems by setting up a model in the original scale, using a *link function* to transform the mean into a linear function of the predictor variables and a *variance function* to allow for variance heterogeneiry in the analysis rather than trying to transform it away. Thus rather than seek a single compromise transformation, GLMs do offer two independent devices, while at the same time the analysis is in the original scale.

One very convenient feature of the linear model is that the predictors enter the model through a single linear function, *only*. More specifically, suppose $\boldsymbol{Y}$ is a response variable and let $x_1, x_2, \ldots, x_p$ be candidate predictors. Let

$$\eta = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p = \boldsymbol{x}^\top\boldsymbol{\beta}$$

be a linear function of the predictors. If the predictors enter the model through η alone, then it is true that *the predictor $\boldsymbol{x}_j$ does not influence the distribution of $\boldsymbol{Y}$ if and only if* $\beta_j = 0$.

The normal linear model is usually written as

$$Y = \boldsymbol{x}^\top\boldsymbol{\beta} + \sigma Z = \eta + \sigma Z \qquad \text{where} \qquad Z \sim \mathrm{N}(0,1)$$

This is clearly the same as describing the model as

$$Y \sim \mathrm{N}\left(\eta = \boldsymbol{x}^\top\boldsymbol{\beta}, \sigma^2\right)$$

Generalized linear models retain the convenient feature that the model depends on the predictors through a single linear function, namely η, only, but relaxes the assumption that the distribution is normal. The assumption is that the distribution of $\boldsymbol{T}$ can be specified as

$$Y \sim f_Y(y; \eta = \boldsymbol{x}^\top\boldsymbol{\beta}, \varphi)$$

where f_Y denotes a generic probability (density) function, belonging to a particular family to be described below in section 4.

The linear function of the predictor variables, η, is called the *linear predictor*.

The precise assumption made is actually a little stronger than this. A generalized linear model for a response variable $\boldsymbol{Y}$ requires that

- The mean of Y depends on the predictor variables through a single linear predictor

$$\mathrm{E}[Y] = \mu = \ell^{-1}(\eta) \qquad \text{or, alternatively,} \qquad \ell(\mu) = \eta = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}$$

  The function, $\ell(.)$ is called the *link function*, and plays a similar rôle to the transformation function.

- The distribution of Y may also involve a constant *scale parameter*, φ.

- The distribution of Y belongs to a particular *family*, to be described below. This will imply that the variance of the distribution has the form

$$\mathrm{var}[Y] = \frac{\varphi}{A} v(\mu)$$

  Here A is a known prior weight and $v(\mu)$ is called the *variance function*, which may depend on the mean, μ, and hence on the linear predictor.

Example 3.1 The following example shows that log-transforming the response and using a GLM with log-link and constant variance can produce dramatically different results. The example is taken from [Ruppert et al.(1989)Ruppert, Cressie & Carroll].

The data gives the weekly percent fat in the milk of a single cow for a period of 35 weeks. The model suggested is of the form

$$\log Y = \beta_0 + \beta_1 w + \beta_2 \log w + \sigma Z \qquad \text{(log-transform)}$$

or

$$Y = \exp(\beta_0 + \beta_1 w + \beta_2 \log w) + \sigma Z \qquad \text{(log-link GLM)}$$

where w is the counter for the week. The latter version is often used in animal science.

```
> Milk <- data.frame(week = 1:35, yield = c(0.31, 0.39, 0.5, 0.58,
        0.59, 0.64, 0.68, 0.66, 0.67, 0.7, 0.72, 0.68, 0.65, 0.64,
        0.57, 0.48, 0.46, 0.45, 0.31, 0.33, 0.36, 0.3, 0.26, 0.34,
        0.29, 0.31, 0.29, 0.2, 0.15, 0.18, 0.11, 0.07, 0.06, 0.01,
        0.01))
> M1 <- lm(log(yield) ~ week + log(week), Milk)
> M2 <- glm(yield ~ week + log(week), quasi(link = "log"), Milk)
> pMilk <- data.frame(week = seq(1, 35, by = 0.1))
> pMilk <- transform(pMilk, pM1 = exp(predict(M1, pMilk)), pM2 = predict(M2,
        pMilk, type = "resp"))
> yl <- range(Milk$yield, pMilk$pM1, pMilk$pM2)
> with(Milk, plot(week, yield, pch = 8, cex = 0.8, xlab = "Week",
        ylab = "Fat yield (kg/day)", ylim = yl, col = "navy"))
> with(pMilk, {
        lines(week, pM1, col = "red")
        lines(week, pM2, col = "green4")
    })
> legend("topright", c("observed", "lm, log-transformed", "glm, log-link"),
        lty = c(NA, "solid", "solid"), pch = c(8, NA, NA), col = c("navy",
            "red", "green4"), cex = 0.8, bty = "n")
```

The result is shown in the left panel of Figure 3. Thr right panel shows the same plot, but with the response on the log scale. As this is the scale in which the log-transformed model has been estimated, it becomes clear why the result appears as it does on the natural scale. The low values on the log scale become very influential points, and on this scale it is the lowest two values which are responsible for the largest residuals.

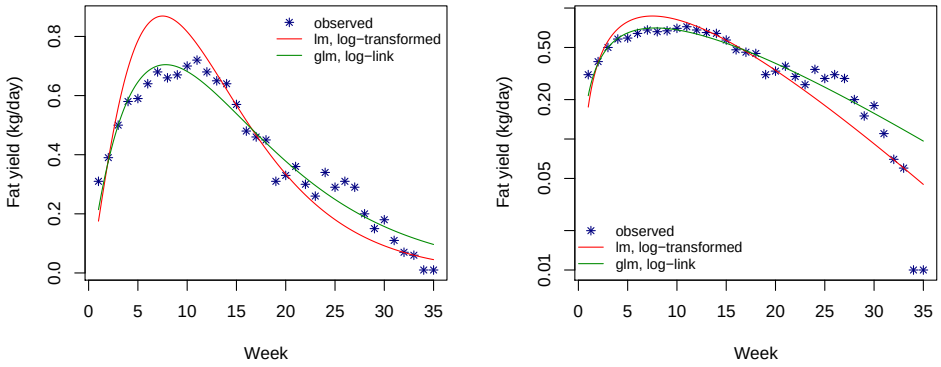

Figure 3: Comparison of log-transform and log-link linear models for the milk fat content data. The left panel shows the repsonse in the natural scale and the right shows it in the log scale.

---

4 The GLM family of distributions

We assume that the distribution to which $\boldsymbol{Y}$ belongs has a probability (density) function that can be written in the form

$$f_Y(y;\eta,\varphi) = \exp\left[\frac{A}{\varphi}\{y\theta - b(\theta)\} + c(y,\varphi)\right]$$

where

- $\boldsymbol{A}$ is a known, positive *prior weight*, not necessarily the same for all observations,
- φ is a constant, positive *scale parameter*, (also called a *dispersion parameter*), possibly, but not necessarily known, and
- $\theta = \theta(\mu)$ depends on the mean, and hence on the linear predictor.

Example 4.1 If $Y \sim \mathrm{N}(\eta,\sigma^2)$, its density function may be written

$$f_Y(y;\eta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{(y-\eta)^2}{2\sigma^2} = \exp\left[\frac{1}{\sigma^2}\left\{y\eta - \frac{\eta^2}{2}\right\} + \frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right]$$

Hence, in this case,

$$A = 1, \quad \varphi = \sigma^2, \quad \theta(\eta) = \eta, \quad b(\theta) = \theta^2/2, \quad \text{and} \quad c(y,\varphi) = \frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)$$

The normal distribution with constant variance is therefore included in the GLM family.

Example 4.2 If Y has a Poisson distribution, that is, $Y \sim \text{Po}(\lambda)$, where $\lambda = \lambda(\eta)$ is some known function of a linear predictor. Then its probability function may be written

$$f_Y(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = \exp\{y\log\lambda - \lambda - \log y!\}$$

Hence $A = 1$, $\varphi = 1$, is a known value in this case, $\theta = \log\lambda(\eta)$, $b(\theta) = \lambda = \exp(\theta)$ and $c(y,\varphi) = -\log y!$. The Poisson distribution belongs to the GLM family.

4.1 Moment generating function and cumulants

The moment generating function for a distribution in the GLM family is

$$M_Y(t) = \text{E}\left[e^{tY}\right] = \int_y e^{ty} f_Y(y;\eta,\varphi)\,dy = \int_y e^{ty}\exp\left[\frac{A}{\varphi}\{y\theta - b(\theta)\} + c(y,\varphi)\right]\,dy$$

where the integration has to be interpreted as over the sample space of Y, (which therefore may be with respect to counting measure if Y is discrete). The integral may be written as

$$M_Y(t) = \exp\left[\frac{A}{\varphi}\left\{b\left(\theta + \frac{t\varphi}{A}\right) - b(\theta)\right\}\right] \times \int_y \exp\left[\frac{A}{\varphi}\left\{y\left(\theta + \frac{t\varphi}{A}\right) - b\left(\theta + \frac{t\varphi}{A}\right)\right\} + c(y,\varphi)\right]\,dy$$

The final integral in this expression is equal to 1, at least for sufficiently small values of t, as it is the integral of a normalised probability density function over the entire sample space. Hence

$$M_Y(t) = \exp\left[\frac{A}{\varphi}\left\{b\left(\theta + \frac{t\varphi}{A}\right) - b(\theta)\right\}\right]$$

The cumulant generating function, $K_Y(t) = \log M_Y(t)$ is then

$$K_Y(t) = \frac{A}{\varphi}\left\{b\left(\theta + \frac{t\varphi}{A}\right) - b(\theta)\right\} = \sum_{j=1}^{\infty}\kappa_j\frac{t^j}{j!}$$

It follows easily that

$$\kappa_1 = \mu = b'(\theta), \quad \kappa_2 = \sigma^2 = \frac{\varphi}{A}b''(\theta) = \frac{\varphi}{A}\frac{d\mu}{d\theta}, \quad \text{and in general} \quad \kappa_r = \left(\frac{\varphi}{A}\right)^{r-1} b^{(r)}(\theta)$$

Notice that this establishes the form of the variance function claimed above:

$$\text{var}[Y] = \frac{\varphi}{A}v(\mu) \quad \text{where the } \textit{variance function} \quad v(\mu) = b''(\theta) = \frac{d\mu}{d\theta}$$

Also, since $A > 0$, $\varphi > 0$ and $\text{var}[Y] > 0$ by definition, it follows that $\frac{d\mu}{d\theta} > 0$. Thus μ must be a monotone increasing, and hence invertable, function of θ. It follows, then, that

$$\frac{d\theta}{d\mu} = \frac{1}{v(\mu)} > 0$$

We use this result in the next section.

4.2 The *natural* link function

For any member of the generalized linear model family one particular link is called the *natural* link because it has properties that make the statistical theory slightly simpler.

Notice that selecting a member of the family implies choosing a particular form for the function $\theta = \theta(\mu)$, and choosing a particular link implies choosing a function, $\ell(.)$, for which $\ell(\mu) = \eta$, or $\mu = \ell^{-1}(\eta)$.

Definition: For any given member of the generalized linear model family the natural link, $\ell_\star$ is that link for which

$$\theta\left(\ell_\star^{-1}(\eta)\right) = \eta$$

identically in η.

Example 4.3 For the $\mathbf{N}(\mu, \sigma^2)$ we saw that $\theta(\mu) = \mu$, which implies that the natural link for this family has $\mu = \eta$, that is $\ell_\star$ is the identity function. This simple case is called the *identity link*.

For the Poisson distributon we say that $\theta(\mu) = \log \mu$, so the natural link is the one for which $\ell_\star^{-1}(\eta) = \exp \eta$, that is $\ell_\star(.) = \log(.)$, the **log** link.

The natural link gives the GLM two important statistical properties, which we state here without proof.

- For the natural link, if the dispersion parameter, φ is *known*, the statistic $\mathbf{X}^\mathsf{T}\mathbf{y}$ is *sufficient* for $\boldsymbol{\beta}$, that is, it determines the likelihood function up to a factor independent of the parameters.
- For the natural link the maximum likelihood estimates satisfy the relationship

  $$\mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{X}^\mathsf{T}\hat{\boldsymbol{\mu}}$$

  This relationship is often described by saying that "the observations and the fitted means have the same *marginal totals*".

In R generalized linear modelling family functions have the natural link as the default, though there is often no particular data analysis reason to prefer it in practice.

5 Estimation

5.1 Some general theory

We begin by stating some definitions and general results we will assume in the later discussion. Proofs may be found in any good intermediate level book on frequentist statistical inference.

Assume $L(\boldsymbol{\gamma}) = \log \mathscr{L}(\boldsymbol{\gamma})$ is a log-likelihood for parameter vector $\boldsymbol{\gamma}^{p\times 1}$.

Definition: The *score vector*, $\mathbf{s}(\boldsymbol{\gamma})$, for $\boldsymbol{\gamma}$ is the vector of first partial derivatives, or *gradient* vector, of $L(\boldsymbol{\gamma})$.

Definition: The *information matrix*, $\mathscr{I}(\boldsymbol{\gamma})$ is the negative of the matrix of second partial derivatives, or *negative Hessian*, of $L(\boldsymbol{\gamma})$.

Definition: The *expected information matrix*, $I(\boldsymbol{\gamma})$ is the expectation of $\mathscr{I}(\boldsymbol{\gamma})$ over the sample space of the observations.

$$\mathbf{s}(\boldsymbol{\gamma}) = \begin{bmatrix} \partial L/\partial\gamma_1 \\ \vdots \\ \partial L/\partial\gamma_p \end{bmatrix} \qquad \mathscr{I}(\boldsymbol{\gamma}) = \begin{bmatrix} -\partial^2 L/\partial\gamma_1^2 & \cdots & -\partial^2 L/\partial\gamma_1\partial\gamma_p \\ \vdots & \ddots & \vdots \\ -\partial^2 L/\partial\gamma_p\partial\gamma_1 & \cdots & -\partial^2 L/\partial\gamma_p^2 \end{bmatrix} \qquad I(\boldsymbol{\gamma}) = \mathrm{E}\left[\mathscr{I}(\boldsymbol{\gamma})\right]$$

It can be shown under mild regularity conditons that

$$\mathrm{E}\left[\mathbf{s}(\boldsymbol{\gamma})\right] = \mathbf{0} \qquad \text{and} \qquad \mathrm{var}\left[\mathbf{s}(\boldsymbol{\gamma})\right] = I(\boldsymbol{\gamma})$$

Further, if $\widehat{\boldsymbol{\gamma}}$ is the MLE, it can usually be shown that, for "large" samples,

$$\mathbf{s}(\boldsymbol{\gamma}) \dot{\sim} \mathrm{N}\left(\mathbf{0}, I(\boldsymbol{\gamma})\right) \qquad \text{and} \qquad \widehat{\boldsymbol{\gamma}} \dot{\sim} \mathrm{N}\left(\boldsymbol{\gamma}, I(\boldsymbol{\gamma})^{-1}\right)$$

Much of statistical inference *in practice* is based on either the latter result and the large sample distribution of the likelihood ratio statistic.

The MLE, in regular cases, occurs at a stationary maximum point of the log-likelihood and most methods for finding it focus on solving the *score equation*, namely

$$\mathbf{s}(\hat{\boldsymbol{\gamma}}) = \mathbf{0}$$

The standard Newton-Raphson method for solving the score equation starts with an inital vector, $\boldsymbol{\gamma}^{(0)}$ and successive approximations are calculated according to the scheme

$$\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)} + \mathscr{I}(\boldsymbol{\gamma}^{(m)})^{-1}\mathbf{s}(\boldsymbol{\gamma}^{(m)})$$

The Fisher modification of this process is to replace the observed information matrix with the expected

$$\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)} + I(\boldsymbol{\gamma}^{(m)})^{-1}\mathbf{s}(\boldsymbol{\gamma}^{(m)})$$

which can simplify the computations without affecting the convergence of the process very much.

There is no guarantee that the process will converge, however, and when it does, there is no guarantee that the point is at the global maximum of the log-likelihood function, but in practice in statistics it is usually the case that convergence is to the MLE and occurs rapidly.

5.2 Estimation of the linear parameters

We *temporarily* assume that $\boldsymbol{\varphi}$ is known and consider estimating the parameters, $\boldsymbol{\beta}$, by maximum likelihood. It is useful to start the discussion with a few formulae for partial derivatives of the log-likelihood.

Write the log-likelihood in the form

$$L = \log \mathscr{L} = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} \left[\frac{A_i}{\varphi} \{ y_i \theta_i - b(\theta_i) \} + c(y_i, \varphi) \right]$$

Using the chain rule

$$\begin{aligned} \frac{\partial L_i}{\partial \beta_j} &= \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ &= \frac{A}{\varphi} \left(y_i - b'(\theta_i) \right) \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \\ &= x_{ij} \times \frac{A}{v(\mu)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \times \frac{y_i - \mu_i}{\partial \mu_i / \partial \eta_i} \times \frac{1}{\varphi} \\ &= x_{ij} \times w_i \times e_i \times 1/\varphi \quad \text{as a definition.} \end{aligned}$$

This will give the score vector for the linear parameters, $\boldsymbol{\beta}$.

Note that the parameter φ occurs only in the final factor. This immediately shows that if we solve the score equations for $\hat{\boldsymbol{\beta}}$ the solution will not depend on φ, and so we may find $\hat{\boldsymbol{\beta}}$ separately first, whether φ is known or not. This will be confirmed later when the φ drops out of the MLE algorithm.

To find the expected information matrix, note that, after a little algebra,

$$\begin{aligned} \mathrm{E}\left[\frac{\partial L_i}{\partial \beta_j} \right] &= 0 \qquad \text{in line with the general result} \\ \operatorname{cov}\left[\frac{\partial L_i}{\partial \beta_j}, \frac{\partial L_i}{\partial \beta_{j'}} \right] = \mathrm{E}\left[\frac{\partial L_i}{\partial \beta_j} \frac{\partial L_i}{\partial \beta_{j'}} \right] &= x_{ij} x_{ij'} w_i \frac{1}{\varphi} \end{aligned}$$

The expected information matrix $\boldsymbol{I}(\boldsymbol{\theta})$, will be obtained by adding these results over all observations.

It is convenient to define quantities in matrix notation. Let $X = (x_{ij})$ be the model matrix and put $W = \operatorname{diag}(w_i)$ as a diagonal matrix of *iterative, or 'working' weights* and $\mathbf{e} = (e_1, \ldots, e_n)^\mathsf{T}$. Then the score vector and information matrix may be written as

$$\begin{aligned} \mathbf{s}(\boldsymbol{\beta}) = \begin{bmatrix} \partial L / \partial \beta_1 \\ \vdots \\ \partial L / \partial \beta_p \end{bmatrix} &= \frac{1}{\varphi} X^\mathsf{T} W \mathbf{e} \\ \boldsymbol{I}(\boldsymbol{\beta}) = \operatorname{var}\left[\mathbf{s}(\boldsymbol{\beta}) \right] &= \frac{1}{\varphi} X^\mathsf{T} W X \end{aligned}$$

Given an initial starting vector, $\boldsymbol{\beta}^{(0)}$, the Fisher modifiet Newton-Raphson process for solvind the score equation defines a series of approximations to the MLE as

$$\begin{aligned} \boldsymbol{\beta}^{(m+1)} &= \boldsymbol{\beta}^{(m)} + I_{(m)}^{-1} \mathbf{s}(\boldsymbol{\beta}^{(m)}) \\ &= \boldsymbol{\beta}^{(m)} + \left(X^\mathsf{T} W_{(m)} X \right)^{-1} X^\mathsf{T} W_{(m)} \mathbf{e}_{(m)} \end{aligned}$$

Notice that the scale parameter, φ, cancels and is not involved at this stage. This equation may also be written as

$$X^\mathsf{T} W_{(m)} X \boldsymbol{\beta}^{(m+1)} = X^\mathsf{T} W_{(m)} \mathbf{y}_{(m)}^\star \qquad \text{where} \qquad \mathbf{y}_{(m)}^\star = X \boldsymbol{\beta}^{(m)} + \mathbf{e}_{(m)}$$

Hence the process may be written as *iterative weighted regression* with iterative weight matrix, $\boldsymbol{W}_{(m)}$ and *working response vector* $\mathbf{y}^{\star}_{(m)} = \boldsymbol{X}\boldsymbol{\beta}^{(m)} + \mathbf{e}_{(m)}$. The components of the vector $\mathbf{e}_{(m)}$ are called the *working residuals*.

This shows again that even if $\boldsymbol{\varphi}$ is not known, it is possible to maximise the likelihood with respect to $\boldsymbol{\beta}$ separately, without it. We could estimate $\boldsymbol{\varphi}$ by maximising the profile likelihood for it, now that we know the MLE for $\hat{\boldsymbol{\beta}}$, but in practice this is not usually done. Instead $\boldsymbol{\varphi}$ is usually estimated by a more informal process, motivated by analogy with the usual estimator for the Normal distribution. We outline this in the next section.

At this stage, though, note that

$$\mathrm{E}\left[\frac{\partial^2 L_i}{\partial\beta_j\partial\varphi}\right] = \mathrm{E}\left[x_{ij} \times \frac{A}{v(\mu)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \times \frac{y_i - \mu_i}{\partial\mu_i/\partial\eta_i} \times \frac{-1}{\varphi^2}\right] = 0$$

This shows that the expected information is block diagonal and hence that $\hat{\boldsymbol{\beta}}$ and $\hat{\varphi}$ are uncorrelated, and hence independent, at least asymptotically.

6 The deviance and estimation of φ

To define the quantity known as the *deviance* of the linear model, we first need to define a model we call the *saturated model*.

Definition: The *saturated model*, denoted by $\boldsymbol{S}$, is a model with n parameters, that is as many parameters as there are observations.

The saturated model is of no interest in data analysis, but note that

- Any genuine model, i.e. one with $p < n$ regression parameters, is a *special case* of the saturated model.
- For the saturated model, a convenient choice for the parameters are the means, μ_i, of the observations themselves. Equating the score vector for μ_i to zero gives the MLE

$$\hat{\mu}_i = y_i \qquad \text{for the saturated model, } \boldsymbol{S}$$

- With n mean parameters, there is no information available to estimate the scale parameter, φ, as well, under $\boldsymbol{S}$.

Assume, temporarily, that the scale parameter, φ, is *known* and that its value is $\varphi = 1$. This assumption is correct for the Binomial and Poisson distributions, but usually not for the Normal, Gamma and Inverse Gaussian distributions.

Now suppose that $\boldsymbol{M}$ is any real model with model matrix $\boldsymbol{X}^{n\times p}$ of rank $p < n$. As noted above, $\boldsymbol{M}$ is a special case of $\boldsymbol{S}$, $\boldsymbol{M} \subset \boldsymbol{S}$, so we can formally consider a *likelihood ratio test statistic* for testing $\boldsymbol{M}$ within $\boldsymbol{S}$:

$$D_M \stackrel{\text{def.}}{=} 2\left(\max_{S} \log\mathscr{L} - \max_{M} \log\mathscr{L}\right)$$

Definition: The quantity D_M is defined as the *deviance* for model $\boldsymbol{M}$.

Likelihood ratio theory would suggest that under the Null hypothesis model, $\boldsymbol{M}$, the deviance, $\boldsymbol{D_M}$, should have an (approximate) chi-squared distribution with $n-p$ degrees of freedom:

$$D_M \dot{\sim} \chi^2(n-p), \qquad \text{if } \boldsymbol{M} \text{ is true and } \varphi = 1$$

This is not strictly supported by Likelihood Ratio theory, as the saturated model does not have a fixed number of parameters. Nevertheless there are situations where it is approximately true, if the initial assumption, $\varphi = 1$, is correct. Hence

$$\mathrm{E}[D_M] \approx n-p \qquad \text{if } \boldsymbol{M} \text{ is true and } \varphi = 1$$

In the case of a Normal model with constant variance, (i.e. the ordinary linear model), it is easy to show that the deviance is just the residual sum of squares:

$$D_M = \mathbf{y}^\top (I - P_X)\mathbf{y} = \mathrm{RSS} \qquad \text{the residual sum of squares, for ordinary linear models}$$

This estimator, the so-called 'reduced' or 'restricted' maximum likelihood estimate, or *REML* estimate for short, can be justified in several ways. For example, it is the MLE got from the *marginal likelihood* for σ^2, which is based on the distribution of RSS itself rather than on the full likelihood.

In this case the scale parameter is the variance, $\varphi = \sigma^2$, so the usual estimate of the scale parameter is

$$s^2 = \frac{\mathrm{RSS}}{n-p} = \frac{D_M}{n-p}$$

Also, if $\boldsymbol{M_0}$ is a sub-model of $\boldsymbol{M}$ with $p_0 < p$ degrees of freedom, the usual test statistic for testing $\boldsymbol{M_0}$ within $\boldsymbol{M}$ is

$$F = \frac{(D_{M_0} - D_M)/(p-p_0)}{D_M/(n-p)}$$

which, if $\boldsymbol{M_0}$ is true, has an $F_{p-p_0,n-p}$ distribution.

In this case it can be shown to be equivalent to the likelihood ratio test, but in the general case it is *not true* that the analogous test is a likelihood ratio test.

Now consider the case of the Binomial or Poisson distribution, where the assumption $\varphi = 1$ is correct. In this case if $\boldsymbol{M_0}$ is a sub-model of $\boldsymbol{M}$, then the likelihood ratio statistic for testing $\boldsymbol{M_0}$ within $\boldsymbol{M}$ is

$$X^2 = D_{M_0} - D_M$$

which, if $\boldsymbol{M_0}$ is true, will have an approximately $\chi^2(p-p_0)$ distribution.

In this case the deviance itself is not needed to estimate a scale parameter, but it is *sometimes* the case that the deviance can be used as a test of fit for the model itself. This result has to be treated with some caution, however. In particular it is usually *not* a reliable test of fit for cases where the data are binary.

6.1 Overdispersion

Very roughly, though, if you were to estimate the scale (or *dispersion*) parameter, as if it were unknown, a natural estimator would be

$$\tilde{\varphi} = \frac{D_M}{n-p}$$

If this estimator is much larger than 1, it is an indication that the model is somewhat doubtful, at least.

Definition: A data set is said to be *overdispersed* with respect to a particular model, if

- The model has known dispersion parameter, $\varphi = 1$,
- For this data set $D_M >> n - p$.

Example 6.1 The Quine data in the `MASS` library give the number of days absent from school in an acedemic year by the students at a particular rural Australian school. The children are further classifies by age group, `Age`, ses, `Sex`, learner status, `Lrn`, and ethnicity, `Eth`. A natural model to consider for this is that the day counts have a Poisson distribution with a mean depending on the subclass to which the child belongs. A quick check, however, shows this to be unrealistic.

```
> quine.1 <- glm(Days ~ Age * Eth * Sex * Lrn, poisson, quine)

> with(quine.1, c(D_M = deviance, "n-p" = df.residual))

     D_M      n-p
1173.899  118.000
```

The deviance is nearly 10 times the residual degrees of freedom. The modelling strategy is clearly not appropriate. It is nevertheles the case that the estimates of the mean parameters are largely unaffected. If standard tests are used, however, ignoring the overdispersion, the resuls will be very misleading. We need to modify the model to take account of the non-Poisson behaviour.

6.2 Uses for the deviance

We now give a short summary on how the deviance is used in generalized linear modelling. There are two separate cases.

Case 1: $\varphi = 1$, known In this case differences of deviance are used as chi-squared tests for sub-models.

$$X^2 = D_{M_0} - D_M \dot{\sim} \chi^2(p - p_0) \quad \text{if } M_0 \text{ is true}$$

The statistic is a true likelihood ratio statistics, and asymptotic likelihood ratio theory generally applies.

The deviance itself, D_M, can sometimes, with caution, be used as an absolute test of fit. Cases where $D_M/(n-p) >> 1$ are usually called *overdispersed* with respect to the model.

The common distributions in this group are the Binomial and Poisson; others include the truncated Binomial, truncated Poisson and Negative Binomial with known shape parameter.

In the rare, but possible, case where φ is known but its value is not 1, the deviance may be re-defined as D_m/φ for the above remarks to apply.

Case 2: φ is unknown In this case the deviance has to be used, (essentially) to estimate φ. There is no overall test of fit based on the deviance, and no concept of "overdispersion".

For the normal distributon the estimate of φ is

$$\tilde{\varphi} = s^2 = \frac{D_M}{n-p}$$

but for other distributions in the group, such as gamma and inverse gaussian, a modified estimator is used, which uses the *Pearson residuals*, as explained below. We also denote this modified estimator by $\tilde{\varphi}$.

Tests for sub-models use an $F-$statistic of the form

$$F = \frac{(D_{M_0} - D_M)/(p - p_0)}{\tilde{\varphi}} \sim F_{p-p_0,n-p} \quad \text{if } M_0 \text{ is true}$$

The main distributions in this group are the normal, gamma and inverse gaussian.

6.3 Residuals

There are four definitions of residuals for generalized linear models routinely available with the GLM model fitting functions in R. These are

Deviance residuals If we think of the deviance as defined as a sum of contributions, one from each observations, each of these contributions has to be positive.

$$D_M = \sum_{i=1}^{n} D_M^{(i)}$$

The deviance residuals are then defined as

$$r_i^D = \sqrt{D_M^{(i)}} \times \text{sign}(y_i - \hat{\mu}_i)$$

Hence the deviance residuals have the properthy that they have the same sign as the differences, $y_i - \hat{\mu}_i$, and their squares sum to the deviance itself.

This is the default definition for residual for GLMs in R and is probably the most useful, simple defintion of residuals to use for most diagnostic purposes. They should in general by regarded *as if* they were normally distributed. However in some cases they will always look far from normal, for example, with strongly discrete data, particularly binary data.

Pearson residuals These are defined as

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)/A_i}}$$

For distributions where φ is unknown, the usual estimate of it is

$$\tilde{\varphi} = \frac{\sum_{i=1}^{n} \left(r_i^P\right)^2}{n-p}$$

This reduces to the "standard" estimate, $D_M/(n-p)$ in the normal case.

Working residuals These come from the iterative algorithm itself at the final stage. They are defined as

$$r_i^W = \frac{y_i - \hat{\mu}_i}{\widehat{\partial \mu_i / \partial \eta_i}} = \hat{e}_i$$

Response residuals Simplest of all. These are defined as the differences

$$r_i^R = y_i - \hat{\mu}_i$$

All four definitions reduce to the same quantity in the case of the normal distribution, but in all other cases some differences exist.

References

[Box & Cox(1964)] Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B* 211–252.

[Ruppert et al.(1989)Ruppert, Cressie & Carroll] Ruppert, D., Cressie, N. & Carroll, R. J. (1989). A transformation/weighting model for estimating Michaelis-menten parameters. *Biometrics* 45 637–656.