Generalized linear models

Douglas Bates

November 01, 2010

Contents

1 Generalized Linear Models

Generalized Linear Models

- When using linear models (LMs) we assume that the response being modeled is on a continuous scale.

- Sometimes we can bend this assumption a bit if the response is an ordinal response with a moderate to large number of levels. For example, the Scottish secondary school test results in the `mlmRev` package are integer values on the scale of 1 to 10 but we analyze them on a continuous scale.

- However, an LM is not suitable for modeling a binary response, an ordinal response with few levels or a response that represents a count. For these we use generalized linear models (GLMs).

- To describe GLMs we return to the representation of the response as an n-dimensional, vector-valued, random variable, $\boldsymbol{\mathcal{Y}}$.

Parts of LMs carried over to GLMs

- Random variables
  - $\boldsymbol{\mathcal{Y}}$ the response variable

- Parameters
  - $\boldsymbol{\beta}$ - fixed-effects coefficients
  - σ - a scale parameter (not always used)
- The linear predictor $\boldsymbol{X\beta}$ where
  - $\boldsymbol{X}$ is the $n \times p$ model matrix for $\boldsymbol{\beta}$

The probability model

- For GLMs we retain some of the properties of the LM probability model

$$\boldsymbol{\mathcal{Y}} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}}, \sigma^2\boldsymbol{I}\right) \text{ where } \boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}} = \boldsymbol{X\beta}$$

  Specifically

  - The distribution of $\boldsymbol{\mathcal{Y}}$ depends on $\boldsymbol{\beta}$ only through the mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}} = \boldsymbol{X\beta}$.
  - Elements of $\boldsymbol{\mathcal{Y}}$ are *independent*. That is, the distribution of $\boldsymbol{\mathcal{Y}}$ is completely specified by the univariate distributions, $\mathcal{Y}_i$, $i = 1, \dots, n$.
  - These univariate, distributions all have the same form. They differ only in their means.
- GLMs differ from LMs in the form of the univariate distributions and in how $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}}$ depends on the linear predictor, $\boldsymbol{X\beta}$.

2 Specific distributions and links

Some choices of univariate distributions

- Typical choices of univariate distributions are:
  - The *Bernoulli* distribution for binary (0/1) data, which has probability mass function

$$p(y|\mu) = \mu^y(1-\mu)^{1-y}, \quad 0 < \mu < 1, \quad y = 0, 1$$

  - Several independent binary responses can be represented as a *binomial* response, but only if all the Bernoulli distributions have the same mean.
  - The *Poisson* distribution for count $(0, 1, \dots)$ data, which has probability mass function

$$p(y|\mu) = e^{-\mu}\frac{\mu^y}{y!}, \quad 0 < \mu, \quad y = 0, 1, 2, \dots$$

- All of these distributions are completely specified by the mean. This is different from the normal (or Gaussian) distribution, which also requires the scale parameter, σ.

The link function, g

- When the univariate distributions have constraints on μ, such as $0 < \mu < 1$ (Bernoulli) or $0 < \mu$ (Poisson), we cannot define the mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}}$, to be equal to the linear predictor, $\boldsymbol{X}\boldsymbol{\beta}$, which is unbounded.
- We choose an invertible, univariate *link function*, g, such that $\eta = g(\mu)$ is unconstrained. The vector-valued link function, $\boldsymbol{g}$, is defined by applying g component-wise.

$$\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\mu}) \quad \text{where} \quad \eta_i = g(\mu_i), \quad i = 1, \ldots, n$$

- We require that g be invertible so that $\mu = g^{-1}(\eta)$ is defined for $-\infty < \eta < \infty$ and is in the appropriate range ($0 < \mu < 1$ for the Bernoulli or $0 < \mu$ for the Poisson). The vector-valued inverse link, $\boldsymbol{g}^{-1}$, is defined component-wise.

"Canonical" link functions

- There are many choices of invertible scalar link functions, g, that we could use for a given set of constraints.
- For the Bernoulli and Poisson distributions, however, one link function arises naturally from the definition of the probability mass function. (The same is true for a few other, related but less frequently used, distributions, such as the gamma distribution.)
- To derive the canonical link, we consider the logarithm of the probability mass function (or, for continuous distributions, the probability density function).
- For distributions in this "exponential" family, the logarithm of the probability mass or density can be written as a sum of terms, some of which depend on the response, y, only and some of which depend on the mean, μ, only. However, only one term depends on **both** y and μ, and this term has the form $y \cdot g(\mu)$, where g is the canonical link.

The canonical link for the Bernoulli distribution

- The logarithm of the probability mass function is

$$\log(p(y|\mu)) = \log(1-\mu) + y \log\left(\frac{\mu}{1-\mu}\right), \; 0 < \mu < 1, \; y = 0, 1.$$

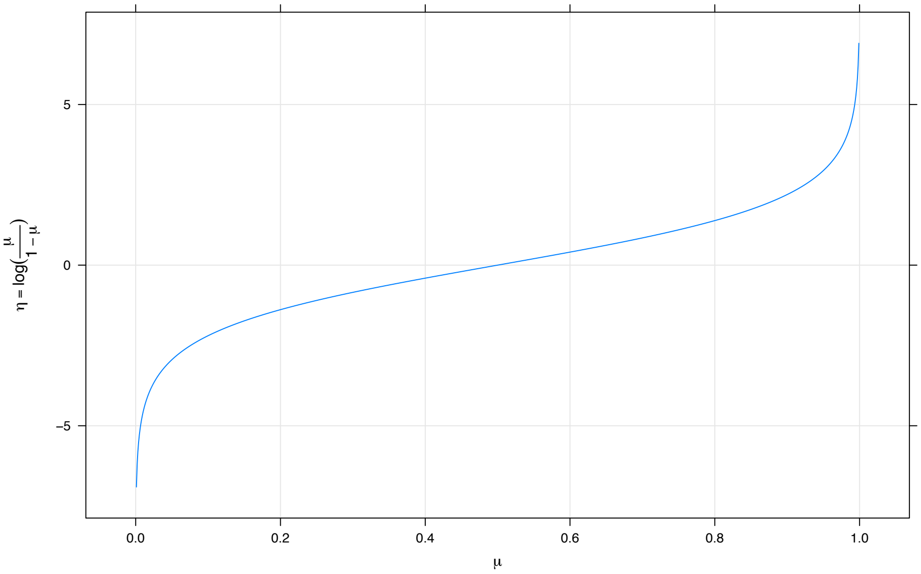
- Thus, the canonical link function is the *logit* link

$$\eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right).$$

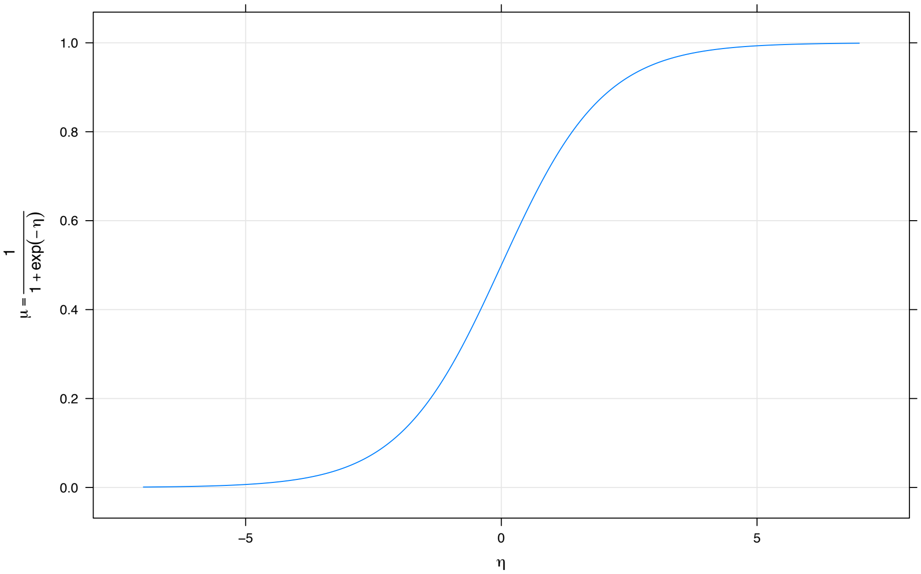
- Because $\mu = P[\mathcal{Y} = 1]$, the quantity $\mu/(1-\mu)$ is the odds ratio (in the range $(0, \infty)$) and g is the logarithm of the odds ratio, sometimes called "log odds".
- The inverse link is

$$\mu = g^{-1}(\eta) = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}}$$

Plot of canonical link for the Bernoulli distribution



Plot of inverse canonical link for the Bernoulli distribution



Evaluating links for the Bernoulli

- As a monotone increasing function the maps $(-\infty, \infty)$ to $(0, 1)$ the inverse link, g^{-1}, is a cumulative distribution function for a continuous random variable. The link, q, is the corresponding quantile function.
- The canonical link is the quantile function for the *logistic* distribution (*R* functions `qlogis` and `plogis`).

- In the past the `"probit"` link was sometimes used instead of the logit link. For this the link is the standard normal quantile and the inverse link is the standard normal cumulative, sometimes written $\Phi(z)$ (*R* functions `qnorm` and `pnorm`).
- As described in *R*'s help page for the `binomial` function

> the `binomial` family (allows) the links `"logit"`, `"probit"`, `"cauchit"`, (corresponding to logistic, normal and Cauchy CDFs respectively) `"log"` and `"cloglog"` (complementary log-log)

The canonical link for the Poisson distribution

- The logarithm of the probability mass is

$$\log(p(y|\mu)) = \log(y!) - \mu + y\log(\mu)$$

- Thus, the canonical link function for the Poisson is the *log* link

$$\eta = g(\mu) = \log(\mu)$$

- The inverse link is

$$\mu = g^{-1}(\eta) = e^{\eta}$$

The canonical link related to the variance

- For the canonical link function, the derivative of its inverse is the variance of the response.
- For the Bernoulli, the canonical link is the logit and the inverse link is $\mu = g^{-1}(\eta) = 1/(1+e^{-\eta})$. Then

$$\frac{d\mu}{d\eta} = \frac{e^{-\eta}}{(1+e^{-\eta})^2} = \frac{1}{1+e^{-\eta}}\frac{e^{-\eta}}{1+e^{-\eta}} = \mu(1-\mu) = \mathrm{Var}(\mathcal{Y})$$

- For the Poisson, the canonical link is the log and the inverse link is $\mu = g^{-1}(\eta) = e^{\eta}$. Then

$$\frac{d\mu}{d\eta} = e^{\eta} = \mu = \mathrm{Var}(\mathcal{Y})$$

3 Estimating parameters

Estimating parameters

- We determine the maximum likelihood estimates, (mle's), of the coefficients, $\boldsymbol{\beta}$, in the linear predictor and, if used, the scale parameter.
- There is no direct algorithm for determining the mle's in a GLM and we must use iterative algorithms. Fortunately, there is a very effective iterative algorithm which is based on *weighted least squares* to update parameter estimates. The weights are inversely proportional to the variance of the response, but the variance depends on the mean which, in turn, depends on the parameters.

- The IRLS (iteratively reweighted least squares) algorithm fixes the weights, determines the parameter values that minimize the weighted sum of squared residuals, then updates the weights and repeats the process until the weights stabilize.

- This algorithm converges very quickly.

- The original description of IRLS from McCullagh and Nelder's book is, I feel, overly complex.

4 Data description and initial exploration

Contraception data

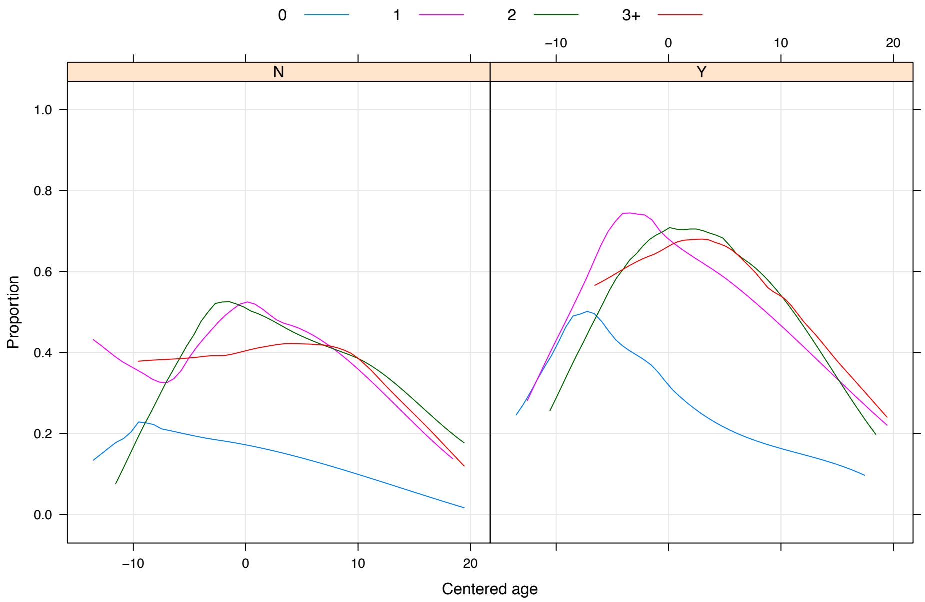
- One of the data sets in the `"mlmRev"` package, derived from data files available on the multilevel modelling web site, is from a fertility survey of women in Bangladesh.

- One of the (binary) responses recorded is whether or not the woman currently uses artificial contraception.

- Covariates included the woman's age (on a centered scale), the number of live children she had, whether she lived in an urban or rural setting, and the district in which she lived.

- Instead of plotting such data as points, we use the 0/1 response to generate scatterplot smoother curves versus age for the different groups.

Contraception use versus age by urban and livch



Comments on the data plot

- These observational data are unbalanced (some districts have only 2 observations, some have nearly 120). They are not longitudinal (no “time” variable).
- Binary responses have low per-observation information content (exactly one bit per observation). Districts with few observations will not contribute strongly to estimates of random effects.
- Within-district plots will be too imprecise so we only examine the global effects in plots.
- The comparisons on the multilevel modelling site are for fits of a model that is linear in `age`, which is clearly inappropriate.
- The form of the curves suggests at least a quadratic in `age`.
- The urban versus rural differences may be additive.
- It appears that the `livch` factor could be dichotomized into “0” versus “1 or more”.

Preliminary model fit

```
Call:
glm(formula = use ~ age + I(age^2) + urban + livch, family = binomial,
    data = Contraception)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4738  -1.0369  -0.6683   1.2401   1.9765

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.9499521  0.1560118  -6.089 1.14e-09
age          0.0045837  0.0089084   0.515    0.607
I(age^2)    -0.0042865  0.0007002  -6.122 9.23e-10
urbanY       0.7680975  0.1061916   7.233 4.72e-13
livch1       0.7831128  0.1569096   4.991 6.01e-07
livch2       0.8549040  0.1783573   4.793 1.64e-06
livch3+      0.8060251  0.1784817   4.516 6.30e-06

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2590.9  on 1933  degrees of freedom
Residual deviance: 2417.7  on 1927  degrees of freedom
AIC: 2431.7

Number of Fisher Scoring iterations: 4
```

Comments on the model fit

- There is a highly significant quadratic term in `age`.
- The linear term in `age` is not significant but we retain it because the `age` scale has been centered at an arbitrary value (which, unfortunately, is not provided with the data).
- The `urban` factor is highly significant (as indicated by the plot).
- Levels of `livch` greater than 0 are significantly different from 0 but may not be different from each other.

5 Model building

Reduced model with dichotomized livch

```
Call:
glm(formula = use ˜ age + I(ageˆ2) + urban + ch, family = binomial,
    data = Contraception)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4544  -1.0371  -0.6682   1.2378   1.9790

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.9429355  0.1497300  -6.298 3.02e-10
age          0.0049675  0.0075889   0.655    0.513
I(ageˆ2)    -0.0043275  0.0006918  -6.255 3.97e-10
urbanY       0.7663616  0.1059463   7.233 4.71e-13
chY          0.8062172  0.1422619   5.667 1.45e-08

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2590.9  on 1933  degrees of freedom
Residual deviance: 2417.9  on 1929  degrees of freedom
AIC: 2427.9

Number of Fisher Scoring iterations: 4
```

Comparing the model fits

- A likelihood ratio test can be used to compare these nested models.

```
> anova(cm2, cm1, test="Chisq")
```

```
Analysis of Deviance Table
Model 1: use ˜ age + I(ageˆ2) + urban + ch
Model 2: use ˜ age + I(ageˆ2) + urban + livch
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      1929     2417.9
2      1927     2417.7  2  0.20525    0.9025
```

- The large p-value indicates that we would not reject `cm2` in favor of `cm1` hence we prefer the more parsimonious `cm2`.
- The plot of the scatterplot smoothers according to live children or none indicates that there may be a difference in the age pattern between these two groups.

6 Conclusions from the example

Conclusions from the example

- Carefully plotting the data is enormously helpful in formulating the model.

- Observational data tend to be unbalanced and have many more covariates than data from a designed experiment. Formulating a model is typically more difficult than in a designed experiment.

- A generalized linear model is fit with the function `glm()` which requires a `family` argument. Typical values are `binomial` or `poisson`. By default, the family will use the canonical link. Other links can be specified as, e.g. `binomial(link="cloglog")`.

- We use likelihood-ratio tests in model building. The z-tests provided in the model summary provide a good indication of the results but should be confirmed by fitting the reduced model and performing a likelihood-ratio test.

A word about overdispersion

- In many application areas using "pseudo" distribution families, such as `quasibinomial` and `quasipoisson`, is a popular and well-accepted technique for accomodating variability that is apparently larger than would be expected from a binomial or a Poisson distribution.

- This amounts to adding an extra parameter, like σ, the common scale parameter in a LM, to the distribution of the response.

- It is possible to form an estimate of such a quantity during the IRLS algorithm but it is an artificial construct. There is no probability distribution with such a parameter.

7 Summary

Summary

- GLMs allow for the distribution of $\mathcal{Y}$ to be other than a Gaussian. A Bernoulli (or, more generally, a binomial) distribution is used to model binary or binomial responses. A Poisson distribution is used to model responses that are counts.

- The mean depends upon the linear predictor, $\boldsymbol{X\beta}$, through the inverse link function, $\boldsymbol{g}^{-1}$.

- The mle's of the parameters are determined through iteratively reweighted least squares (IRLS).

- "Wald-type" tests on the coefficients are displayed in the summary but to be more confident of the results we should fit both models and use a comparative anova with the optional argument, `test="Chisq"`.