Generalized Linear Models

Rosalba Radice

Analytics Methods for Business

Review on Linear Model

- In a linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

  the response y_i, $i = 1, \ldots, n$ is modelled by a linear function of explanatory variables x_j, $j = 1, \ldots, p$ plus an error term.

- We assume that the errors ϵ_i are independent and identically distributed such that

$$\mathrm{E}[\epsilon_i] = 0 \quad \text{and } \mathrm{Var}(\epsilon_i) = \sigma^2.$$

- Typically we assume

$$\epsilon_i \sim N(0, \sigma^2)$$

  as a basis for inference, e.g. t-tests on parameters.

Restrictions of Linear Models

Although a very useful framework, there are some situations where general linear models are not appropriate:

- The range of y is restricted (e.g. binary, count, positive)
- The variance of y depends on the mean

Generalized linear models extend the general linear model framework to address both of these issues.

Generalized Linear Models

A generalized linear model is made up of a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

and two functions:

- a link function that describes how the mean, $\mathrm{E}(y_i) = \mu_i$, depends on the linear predictor
$$g(\mu_i) = \eta_i$$
- a variance function that describes how the variance, $\mathrm{var}(y_i)$, depends on the mean
$$\mathrm{Var}(y_i) = \phi V(\mu_i)$$
where the dispersion parameter ϕ is a constant.

Normal Linear Model as a Special Case

For the linear model with $y_i \sim N(\mu_i, \sigma^2)$ we have the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

the link function

$$g(\mu_i) = \mu_i$$

and the variance function

$$V(\mu_i) = 1.$$

Binomial Data

Suppose

$$y_i \sim Binomial\left(n_i, \frac{\mu_i}{n_i}\right)$$

and we wish to model the proportions y_i/n_i. Then

$$\mathrm{E}(y_i) = \mu_i \quad \text{and} \quad \mathrm{Var}(y_i) = \mu_i\left(1 - \frac{\mu_i}{n_i}\right).$$

The variance function is

$$V(\mu_i) = \mu_i\left(1 - \frac{\mu_i}{n_i}\right).$$

Our link function must map from $(0, 1) \to (-\infty, \infty)$. A common choice is

$$g(\mu_i) = \log\left(\frac{\mu_i}{n_i - \mu_i}\right).$$

Positive Continuous Data

A gamma GLM is of the form

$$y_i \sim Gamma(\mu_i, \phi),$$

with

$$\mathsf{E}(y_i) = \mu_i \quad \text{and} \quad \mathsf{Var}(y_i) = \mu_i^2 \phi.$$

The variance function is

$$V(\mu_i) = \mu_i^2.$$

Our link function must map from $(0, \infty) \to (-\infty, \infty)$. A common choice is

$$g(\mu_i) = \log(\mu_i).$$

Count Data

Suppose

$$y_i \sim Poisson(\mu_i)$$

then

$$\mathrm{E}(y_i) = \mu_i \quad \text{and} \quad \mathrm{Var}(y_i) = \mu_i.$$

So our variance function is

$$V(\mu_i) = \mu_i.$$

Our link function must map from $(0, \infty) \to (-\infty, \infty)$. A natural choice is

$$g(\mu_i) = \log(\mu_i).$$

Transformation versus GLM

In some situations a response variable can be transformed to improve linearity and homogeneity of the variance so that a general linear model can be applied. This approach has some drawbacks:

- The response variable has changed.
- Transformation must simultaneously improve linearity and homogeneity of variance.

Transformation versus GLM - continued

For example, a common remedy for the variance increasing with the mean is to apply the log transform, e.g.

$$\begin{aligned} \log(y_i) &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \Rightarrow \mathrm{E}(\log y_i) &= \beta_0 + \beta_1 x_i \end{aligned}$$

This is a linear model for the mean of $\log y$ which may not always be appropriate. E.g. if y is income perhaps we are really interested in the mean income of population subgroups, in which case it would be better to model $\mathrm{E}(y)$ using a glm:

$$\log \mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$$

with a gamma distribution and a log link.

Exponential Family

Most of the commonly used statistical distributions, e.g. normal, binomial, gamma and Poisson, are members of the exponential family of distributions whose densities can be written in the form

$$f(y;\theta,\phi) = \exp\{[y\theta - b(\theta)]/\phi + c(y,\phi)\}$$

for functions $b(\cdot)$, $c(\cdot)$ and parameters θ (canonical parameter) and ϕ (dispersion parameter).
It can be shown that

$$\mathrm{E}(Y) = b'(\theta) \quad \text{and} \quad \mathrm{Var}(Y) = \phi b''(\theta).$$

Link Function

- For any particular problem, it is possible that there may be several plausible candidates for the link function. An interesting challenge is to propose the most suitable candidate for the problem at hand.
- However, for any distribution, we can associate with it a default, or *canonical* link function.
- The canonical link is defined to be that function $g(\cdot)$ for which

$$g(\mu) = \theta = \eta.$$

- Canonical links lead to desirable statistical properties of the glm hence tend to be used by default.

Estimation of the Model Parameters

- An iterative algorithm can be used to estimate the parameters using maximum likelihood. The log-likelihood is

$$\ell = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\} = \sum_{i=1}^{n} \ell_i.$$

- The maximum likelihood estimates are obtained by solving the score equations

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0$$

  for parameters β_j.
- In general the score equations are non-linear in the β_j, and require numerical iterative techniques for their solution.
- $\hat{\beta}_j$ have the usual properties of maximum likelihood estimators.
- ϕ is usually estimated by method of moments.

Modelling Count Data

Examples of count data include

- Number of household burglaries in a city in a given year.
- Number of customers served by a saleperson in a given month.
- Number of train accidents in a given year.

In such situations, the counts can be assumed to follow a Poisson distribution, say

$$y_i \sim Poisson(\mu_i),$$

where

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

Modelling Count Data - continued

In many cases we are making comparisons across observation units i with different levels of exposure to the event and hence the measure of interest is the rate of occurrence:

- Number of household burglaries per 10,000 households in city i in a given year.
- Number of customers served per hour by salesperson i in a given month.
- Number of train accidents per billion train-kilometers in year i.

Modelling Count Data - continued

Since the counts are Poisson distributed, we would like to use a GLM to model the expected rate, μ_i/t_i, where t_i is the exposure for unit i.
The model would be

$$\log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip},$$

that is:

$$\log\left(\mu_i\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \log(t_i),$$

i.e. Poisson GLM with the canonical log link.

The standardizing term $\log(t_i)$ is an example of an offset: a term with a fixed coefficient of 1.

Ships Data Example

- The `ships` from the `MASS` package concern a type of damage caused by waves to the forward section of cargo-carrying vessels.

- The variables are number of damage incidents (`incidents`), aggregate months of service (`service`), period of operation (`period`): 1960-74, 75-79, year of construction (`year`): 1960-64, 65-69, 70-74, 75-79, and type (`type`): "A" to "E".

- Let us consider a GLM Poisson model with log link including all the variables:

$$\log(\mu) = \beta_0 + \beta_1 \texttt{typeB} + \beta_2 \texttt{typeC} + \beta_3 \texttt{typeD} + \beta_4 \texttt{typeE} + \beta_5 \texttt{year65} + \beta_6 \texttt{year70} + \beta_7 \texttt{year75} + \beta_8 \texttt{period75} + \log(\texttt{service})$$

- The deviance is somewhat larger than the degrees of freedom which indicates lack of fit (deviance $= 38.695$ on 25 degrees of freedom).

Overdispersion

Lack of fit may be due to inadequate specification of the model, but another possibility when modelling discrete data is overdispersion.
Under the Poisson model, we have a fixed mean-variance relationship

$$\text{Var}(y_i) = \text{E}(y_i) = (\mu_i).$$

Overdispersion occurs when

$$\text{Var}(y_i) > \mu_i.$$

This may occur due to correlated responses or variability between observational units.

We can adjust for over-dispersion by estimating a dispersion parameter ϕ

$$\text{Var}(y_i) = \phi V(\mu_i).$$

Models for Overdispersion

- In the `ships` data, it is likely that there is inter-ship variability in accident-proneness.
- We can switch to a quasi-likelihood estimation using the corresponding quasi-family.
- The dispersion parameter is estimated as 1.69, much larger than the value of 1 assumed under the Poisson model.
- Another possible remedy is to consider a more flexible distribution that does not impose equality of mean and variance: the negative binomial.

- The variance of the negative binomial distribution is given by

$$\text{Var}(y) = \mu + \frac{\mu^2}{\tau},$$

  where μ is the mean and τ the shape parameter.

- The Poisson distribution with parameter μ arises for $\tau \to \infty$.

Interpretation of the Model Parameters

Consider ships of type C and E. We have

$$\log(\mu_i^E) - \log(\mu_i^C) = \log(\texttt{service}_i^E) - \log(\texttt{service}_i^C) + \beta_4 - \beta_2$$

and

$$\beta_4 - \beta_2 = \log\left(\frac{\mu_i^E}{\texttt{service}_i^E}\right) - \log\left(\frac{\mu_i^C}{\texttt{service}_i^C}\right) = \log\left(\frac{r_i^E}{r_i^C}\right).$$

So $\exp(\beta_4 - \beta_2)$ is the ratio of the rates (expected number of damages per month in service).

Interpretation of the Model Parameters - continued

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	-6.40590	0.21744	-29.460	< 0.00000
`typeB`	-0.54334	0.17759	-3.060	0.00222
`typeC`	-0.68740	0.32904	-2.089	0.03670
`typeD`	-0.07596	0.29058	-0.261	0.79377
`typeE`	0.32558	0.23588	1.380	0.16750
`year65`	0.69714	0.14964	4.659	< 0.00000
`year70`	0.81843	0.16977	4.821	< 0.00000
`year75`	0.45343	0.23317	1.945	0.05182
`period75`	0.38447	0.11827	3.251	0.00115

- Types B and C have the lowest risk, E the highest (as compared to A). The rate for E is $\exp(0.33 - (-0.69)) = 2.75$ times that for C.
- The incident rate increased by a factor of $\exp(0.38) = 1.47$ after 1974.
- The ships built between 1960 and 1964 seem to be the safest, with ships built between 1965 and 1974 having the highest risk.

Modelling Positive Continuous Data

The gamma distribution can be used in a range of disciplines. Examples of events that may be modeled by gamma distribution include:

- The amount of rainfall accumulated in a reservoir.
- The size of loan defaults and insurance claim cost.
- The flow of items through manufacturing and distribution processes.
- The load on web servers.

Modelling Positive Continuous Data - continued

A gamma GLM is of the form

$$y_i \sim Gamma(\mu_i, \phi)$$

with

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

The canonical link for the gamma distribution is the inverse function. Since parameters from a model with inverse link are difficult to interpret, the log link is usually regarded as more useful.

Personal Injury Insurance Data Example

- This data set contains information on 22,036 settled personal injury insurance claims.
- These claims arose from accidents occurring from July 1989 through to January 1999. Claims settled with zero payment are not included.
- The variables considered are: claim size (`total`), operational time (`op_time`) and legal representation (`legrep`).
- The linear predictor for this model is

$$\log(\mu) = \beta_0 + \beta_1 \texttt{op\_time} + \beta_2 \texttt{legrep}.$$

Interpretation of the Model Parameters

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	8.2118447	0.0329095	249.528	< 0.0000
`op_time`	0.0383149	0.0006311	60.707	< 0.0000
`legrepYes`	0.4667863	0.0424613	10.993	< 0.0000

- $\exp(0.467) = 1.59$ means that claims with legal representation are 1.6 times more likely than claims with no legal representation. That is, we would predict that there are 60% more claims if there is legal representation than there are claims if there is no legal representation.

- If we increase of one unit the operational time we would expect that the predicted claim size would be multiplied by $\exp(0.04) = 1.04$, an increase of 4%.