Generalized Linear Models

Statistics 220

Spring 2005

Generalized Linear Models

For many problems, standard linear regression approaches don't work. Sometimes, transformations will help, but not always. Generalized Linear Models are an extension to linear models which allow for regression in more complex situations.

As before let y be the response variable and X be the predictor variables. We want to determine $p(y|X)$, which will depend on some parameters β and ϕ.

- Non-linearity, multiplicative effects and errors
- Bounded responses
- Discrete responses

Generalized linear model involve the following 4 pieces.

1. Linear predictor: $\eta = X\beta$

2. Link function $g(\cdot)$: Relates the linear predictor to the mean of the outcome variable

$$g(\mu) = \eta = X\beta \qquad \mu = g^{-1}(\eta) = g^{-1}(X\beta)$$

3. Distribution: What is the distribution of the response variable y. These are usually a member of the exponential family which includes, normal, lognormal, poisson, binomial, gamma, hypergeometric.

4. Dispersion parameter ϕ: Some distributions have an additional parameter dealing with the the spread of the distribution. The form of this usually depends on the relationship between the mean and the variance. With some distributions, this is fixed (e.g. Poisson or binomial), while with others it is an additional parameter to the modelled and estimated (e.g. normal or gamma).

Normal linear regression is a special case of a generalized linear model were

1. Linear predictor: $\eta = X\beta$
2. Link function: $g(\mu) = \mu = X\beta$ (Identity Link)
3. Distribution: $y_i|x_i, \beta, \sigma^2 \sim N(x_i^T\beta, \sigma^2)$
4. Dispersion parameter: σ^2

We have seen other GLIMs in class this term.

- Poisson regression:

  The usually form for Poisson regression uses the log link

$$\log \mu = X\beta \qquad \mu = \exp(X\beta)$$

  Another example used the identity link

$$\mu = X\beta$$

  This is less common as the mean of a Poisson random variable must be positive which the identity link doesn't guarantee. The log link however does, which is one reason it is very popular.

  As mentioned earlier the dispersion parameter $\phi = 1$ is fixed as $\mathrm{Var}(y) = \mu$ for the Poisson distribution.

- Logistic regression:

  This form is slightly different as we work with the mean of the sample proportion $\frac{y_i}{n_i}$ instead the mean of y_i.

  Logistic regression is based on $y_i|x_i \sim Bin(n_i, \mu_i)$ where μ_i is a function of x_i. The link function is

$$g(\mu) = \log \frac{\mu}{1-\mu}$$

  i.e. the log odds ratio.

  The inverse link function gives

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

Thus the likelihood is

$$p(y|\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} \left(\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right)^{y_i} \left(\frac{1}{1+e^{X_i\beta}}\right)^{n_i-y_i}$$

The dispersion parameter $\phi = 1$

Link Functions

Changing the link functions allows for different relationships between the response and predictor variables. The choice of link function $g(\cdot)$ should be made so that the relationship between the transformed mean and the predictor variables is linear.

Note transforming the mean via the link function is different from transforming the data

For example consider the two models

1. $\log y_i | X_i, \beta \sim N(X_i\beta, \sigma^2)$ or equivalently $y_i | X_i, \beta \sim logN(X_i\beta, \sigma^2)$

$$E[y_i | X_i, \beta] = \exp\left(X_i\beta + \frac{\sigma^2}{2}\right)$$

and

$$\mathrm{Var}(y_i | X_i, \beta) = \exp(2(X_i\beta + \sigma^2))(\exp(\sigma^2) - 1)$$

2. $y_i|X_i, \beta \sim N(\mu_i, \sigma^2)$ where $\log \mu_i = X_i\beta, \mu_i = \exp(X_i\beta)$ (normal model with log link)

The first model has a different mean and the variability depends on X where as the variability in the second model does not depend on X.

When choosing a link function, you often need to consider the plausible values of the mean of the distribution.

For example, with binomial data, the success probability must be in [0,1]. However $X\beta$ can take values on $(-\infty, \infty)$.

Thus you can get into trouble with binomial data with the model $\mu = X\beta$ (identity link).

Possible choices include

- Logit link:

$$g(\mu) = \log \frac{\mu}{1-\mu}$$

- Probit link:

$$g(\mu) = \Phi^{-1}(\mu) \qquad (\text{Standard Normal Inverse CDF})$$

- Complementary Log-Log link

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

All of these happen to be quantile functions for different distributions.

Thus the inverse link functions are CDFs

- Logit link:

$$g^{-1}(\eta) = \frac{e^{\eta}}{1 + e^{\eta}} \qquad (\text{Standard Logistic})$$

- Probit link:

$$g^{-1}(\eta) = \Phi(\eta) \qquad (N(0, 1))$$

- Complementary Log-Log link:

$$g^{-1}(\eta) = e^{-e^{\eta}} \qquad (\text{Gumbel})$$

Thus in this case any distribution defined on $(-\infty, \infty)$ could be the basis for a link function, but these are the popular ones. One other choice that is used are based on t_ν distributions as they have some robustness properties.

Note that a link function doesn't have to have the property of mapping the range of the mean to $(-\infty, \infty)$.

We've seen the identity link $(g(\mu) = \mu)$ in Poisson regression and it is also used in binomial problem.

In the binomial case, it can be reasonable if the success probabilities lie in the range (0.2, 0.8).

Similarly, an inverse link function doesn't have to have to map $X\beta$ back to the whole range of the mean for a distribution.

For example, the log link will only give positive means ($\mu = e^\eta$). This can be an useful model with normal data, even though in general a normal mean can take any value.

Common Link Functions

The following are common link function choices for different distributions

- Normal
  - Identity: $g(\mu) = \mu$
  - Log: $g(\mu) = \log \mu$
  - Inverse: $g(\mu) = \frac{1}{\mu}$
- Binomial
  - Logit: $g(\mu) = \log \frac{\mu}{1-\mu}$
  - Probit: $g(\mu) = \Phi^{-1}(\mu)$
  - Complementary Log-Log link: $g(\mu) = \log(-\log(\mu))$
  - Log: $g(\mu) = \log \mu$

- Poisson
  - Log: $g(\mu) = \log \mu$
  - Identity: $g(\mu) = \mu$
  - Square root: $g(\mu) = \sqrt{\mu}$
- Gamma
  - Inverse: $g(\mu) = \frac{1}{\mu}$
  - Log: $g(\mu) = \log \mu$
  - Identity: $g(\mu) = \mu$
- Inv-Normal
  - Inverse squared: $g(\mu) = \frac{1}{\mu^2}$
  - Inverse: $g(\mu) = \frac{1}{\mu}$
  - Log: $g(\mu) = \log \mu$
  - Identity: $g(\mu) = \mu$

The first link function mentioned for each distribution is the canonical link which is based on the writing the density of each distribution in the exponential family form.

$$p(y|\theta) = f(y)g(\theta)\exp\left(\phi(\theta)^T u(y)\right)$$

Dispersion Parameter

So far we have only discussed the mean function. However we also need to consider the variability of the data as well. For any distribution, we can consider the variance to be a function of the mean ($V(\mu)$) and a dispersion parameter (ϕ)

$$\text{Var}(y) = \phi V(\mu)$$

The variance functions and dispersion parameters for the common distributions are

Distribution	$N(\mu, \sigma^2)$	$Pois(\mu)$	$Bin(n, \mu)$	$Gamma(\alpha, \nu)$
$V(\mu)$	1	μ	$\mu(1-\mu)$	μ
ϕ	σ^2	1	$\frac{1}{n}$	$\frac{1}{\nu}$

Note for the Gamma distribution, the form of these can depend on how the distribution is parameterized. (McCullagh and Nelder have different formulas due to this.)

So when building models we need models for dealing with the dispersion in the data. Exactly how you want to do this will depend on the problem.

Overdispersion

Often data will have more variability than might be expected.

For example, consider Poisson like data and consider a subset of data which has the same levels of the predictor variables (call it $y_1, y_2, \ldots, y_m$).

If the data is Poisson, the sample variance should be approximately the sample mean

$$s_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2 \approx \bar{y}$$

If $s_m^2 > \bar{y}$, this suggests that there is more variability than can be explained by the explanatory variables.

This extra variability can be handled in a number of ways.

One approach is to add in some additional variability into the means.

$$\begin{aligned} y_i|\mu_i &\overset{ind}{\sim} Pois(\mu_i) \\ \mu_i|X_i,\beta,\sigma^2 &\overset{ind}{\sim} N(X_i\beta,\sigma^2) \end{aligned}$$

In this approach every observation with the same level of the explanatory variables will have a different mean, which will lead to more variability in the ys.

$$\begin{aligned} \mathrm{Var}(y_i) &= E[\mathrm{Var}(y_i|\mu_i)] + \mathrm{Var}(E[y_i|\mu_i]) \\ &= E[\mu_i] + \mathrm{Var}(\mu_i) \\ &= X_i\beta + \sigma^2 \geq X_i\beta = E[y_i] \end{aligned}$$

Note that normally you probably model $\log \mu_i \sim N(X_i\beta,\sigma^2)$, but showing the math was easier with the identity link instead of the log link.

Bayesian Approach to Generalized Linear Models

So far there really hasn't been any Bayes in the discussion so far. Lets now look at the complete model with priors added assuming no overdispersion.

$$\begin{aligned} y_i|\mu_i,\phi &\overset{ind}{\sim} p(y_i|\mu_i,\phi) \\ g(\mu_i) &= X_i\beta \qquad\qquad \text{or } \mu_i = g^{-1}(X_i\beta) \\ \beta &\sim p(\beta) \\ \phi &\sim p(\phi) \end{aligned}$$

The likelihood piece of the model may not correspond to the usual parameterization of the model, and thus the usual parameters will have to be calculated from μ_i and ϕ.

The model for β is general and could have an additional hierarchical structure following the ideas from chapter 15.

For example, if two of the explanatory variable are categorical factors, the β will need to have an ANOVA like structure to account for these.

In addition, the dispersion parameter in this framework is the same for all observations, but it could be allowed to vary from observation to observation by having it depend on X_i. The classical approach to GLIMs usually doesn't do this.

To allow for overdispersion a distribution can be put on μ_i, or what is usually easier on $g(\mu_i)$

$$g(\mu_i) \sim p(\mu_i|X_i\beta)$$

Note that $g(\mu_i) = X_i\beta$ is a special case of this.

Choice of Priors

There are a number of approaches used for putting priors on the regression and dispersion parameters. The usual approach factorizes the prior as

$$p(\beta, \phi) = p(\beta|\phi)p(\phi)$$

Often β and ϕ will have independent priors.

Then the prior on β is commonly done in one of the three ways

- Noninformative: A flat improper prior can be put on β, which yields the classical analysis of the GLIM. Thus the MLE is the posterior mode. This can be determined with standard GLIM software (such as `gls` in `R`). Approximate inference can be obtained by a normal approximation to the likelihood.

- Conjugate: A conjugate prior can be implemented by n_0 idealized observations y_0 with covariate matrix X_0. Inference is carried out by performing analysis on the augmented response variable $\binom{y}{y_0}$ with augmented covariate matrix $\binom{X}{X_0}$ and a flat prior on β. Then inference is performed as in the noninformative prior case.

- Non-conjugate: The most common approach is to express prior information directly on β. A common approach is to use

$$\beta \sim N(\beta_0, \Sigma_\beta)$$

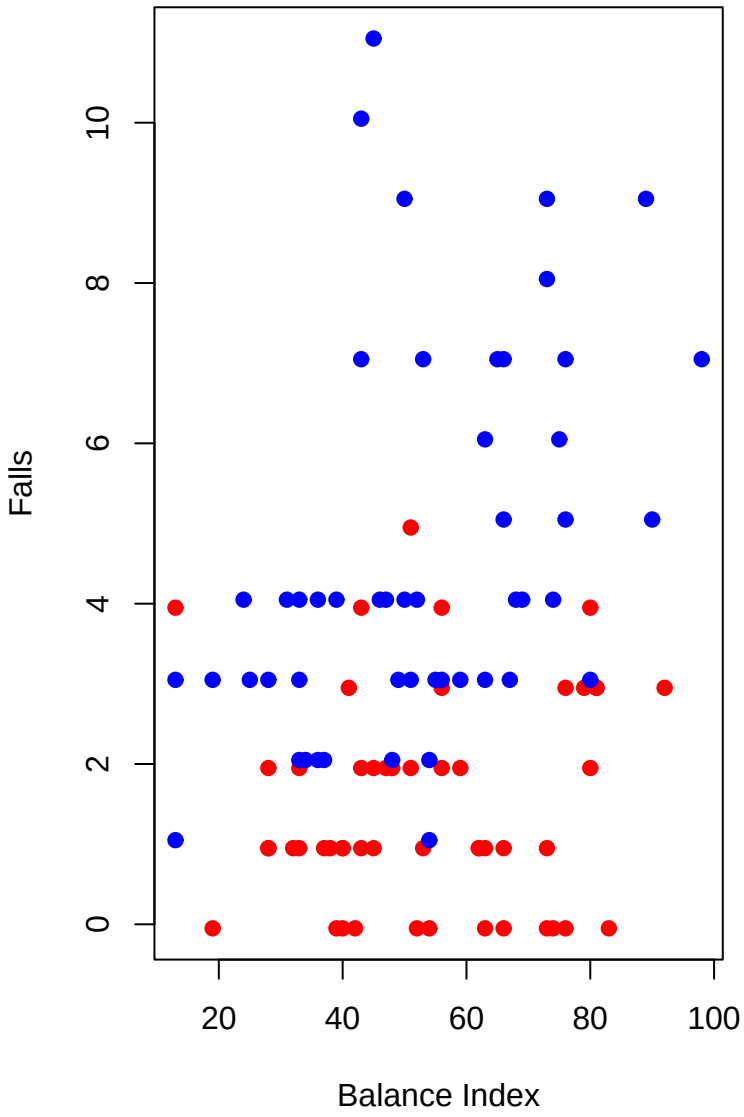
  While the normal is common, other distributions can be used. The normal is convenient when approximate inference based on normal approximations are used. It also has the advantage that it fits into the way people often think of uncertainty, in terms of means and standard deviations (or variances).

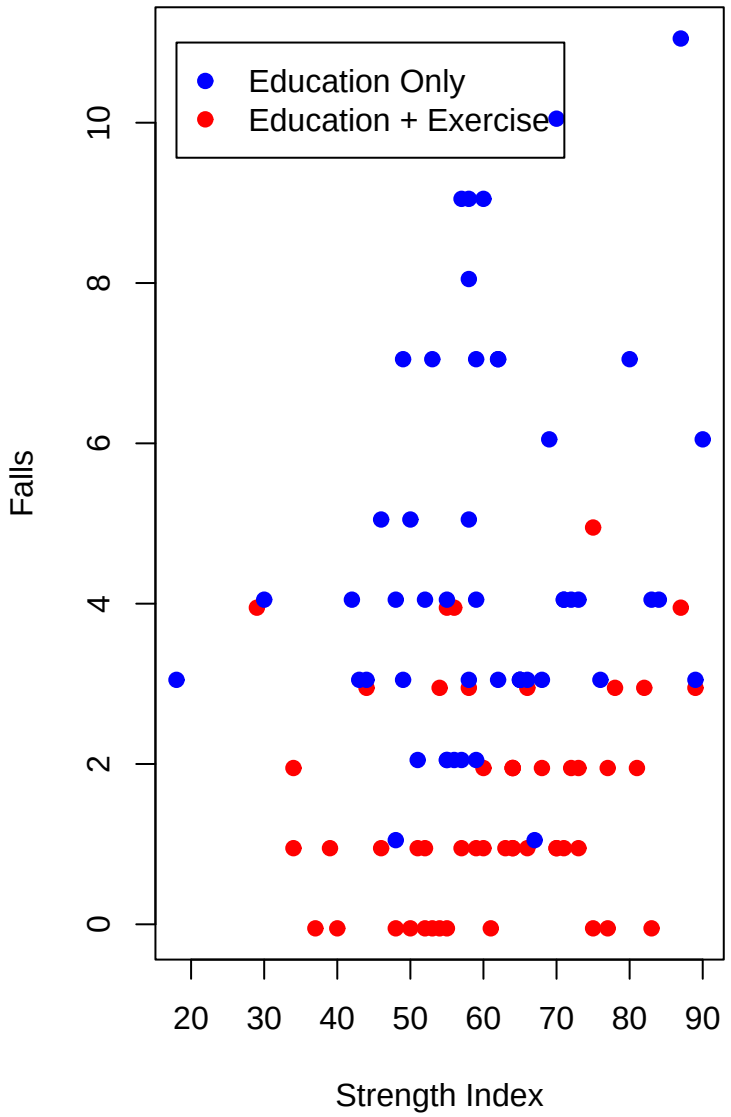
Example

Geriatric Study to Reduce Falls

100 subject were studied to investigate two treatments to which is better to reduce falls.

- y: number of falls during 6 months of study (self-reported)
- x_1: Treatment - 0 = education only, 1 = education + aerobic exercise
- x_2: Gender - 0 = female, 1 = male
- x_3: Balance Index (bigger is better)
- x_4: Strength Index (bigger is better)


Falls
0
2
4
6
8
10
20
40
60
80
100
Balance Index


Education Only
Education + Exercise
Falls
0
2
4
6
8
10
20
30
40
50
60
70
80
90
Strength Index

1. Overdispersion

$$\begin{aligned} y_i|\mu_i &\overset{ind}{\sim} Pois(\mu_i) \\ \log\mu_i|\beta,\sigma^2 &\overset{ind}{\sim} N(X_i\beta,\sigma^2) \\ \beta|\sigma^2_\beta &\sim N(0,\sigma^2_\beta I) \\ \sigma^2_\beta &\sim \text{Inv}-\text{Gamma}(0.001,0.001) \\ \sigma^2 &\sim U(0,1000) \end{aligned}$$

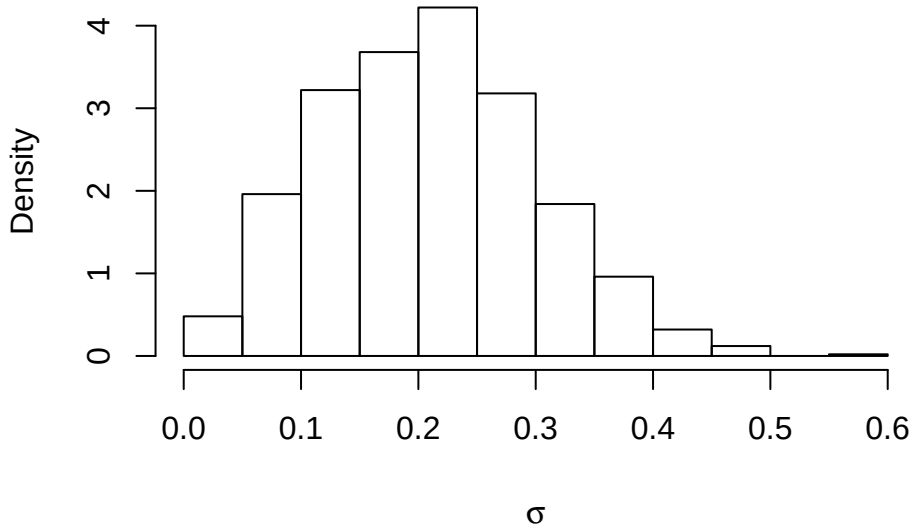
2. No Overdispersion

$$\begin{aligned} y_i|\mu_i &\overset{ind}{\sim} Pois(\mu_i) \\ \log\mu_i &= X_i\beta \\ \beta|\sigma^2_\beta &\sim N(0,\sigma^2_\beta I) \\ \sigma^2_\beta &\sim \text{Inv}-\text{Gamma}(0.001,0.001) \end{aligned}$$

First lets examine whether we need the overdispersion parameter σ^2

First lets look at its posterior distribution

$$\begin{aligned} E[\sigma|y] &= 0.21 \\ \mathrm{SD}(\sigma|y) &= 0.09 \end{aligned}$$



This suggests that if there is any overdispersion, it must be small.

In addition lets look at the DICs

Model	DIC	p_D
Overdispersion	402.9	41
No Overdispersion	378.9	6.1

This implies we are getting a better fit without the overdispersion.

Now lets examine whether the treatment has any effect based on the No Overdispersion model

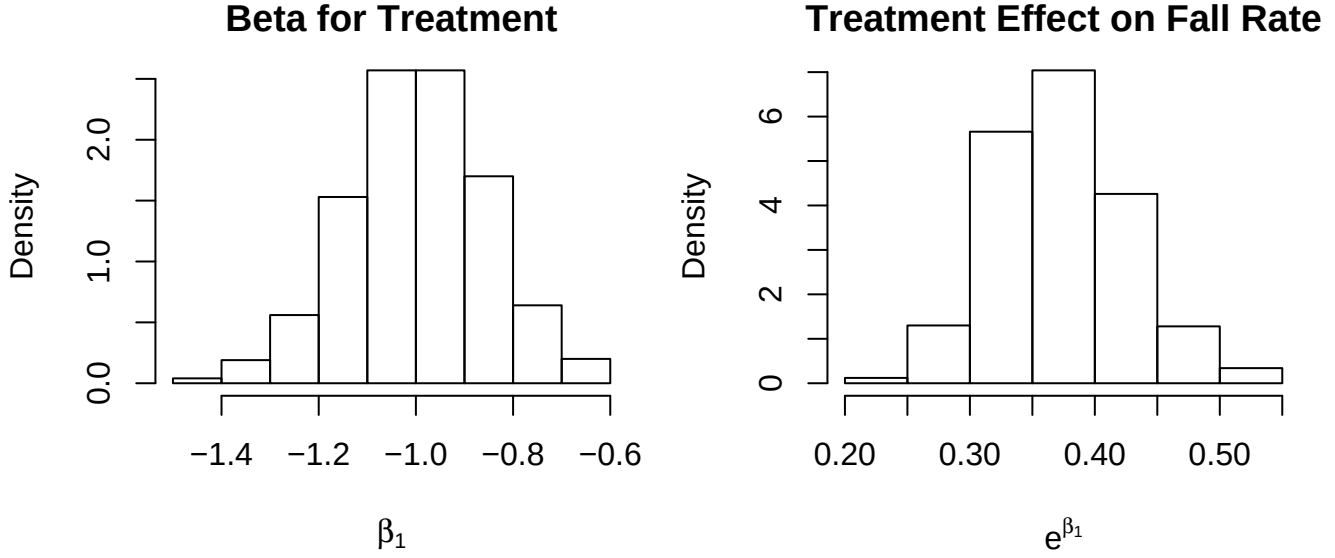
Parameter	$E[\beta_j\|y]$	$\mathrm{SD}(\beta_j\|y)$
Intercept	0.443	0.338
Treatment	-1.010	0.139
Gender	-0.050	0.122
Balance	0.010	0.003
Strength	0.009	0.004

There is strong evidence that the aerobic exercise helps as $E[\beta_1|y] < 0$ ($P[\beta_1 < 0|y] = 1$).

As we are using a log link here

$$\mu_i = e^{\beta_1 x_{1i}} \exp(\beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i})$$

If everything else is kept the same, exercise should lower the fall rate by a factor of $0.36 = e^{-1.010}$. So the exercise should lower the fall rate to about a third of what it would be with education only.



Based on the posterior distribution of e^{β_1} we have strong evidence that the fall rate should at least halve.

Parameter	$E[\beta_j \mid y]$	$\mathrm{SD}(\beta_j \mid y)$
Intercept	0.443	0.338
Treatment	-1.010	0.139
Gender	-0.050	0.122
Balance	0.010	0.003
Strength	0.009	0.004

It appears that there is no significant gender effect as β_2 appears to be close to zero.

There is at first look a slight surprising result for β_3 and β_4. At first thought you might think that people that have better balance and strength might fall less. However these parameter estimates suggest that they fall more. One possible explanation is that the stronger people are more active, meaning they might have more opportunity to fall.