Generalized Linear Models Introduction

Statistics 135

Autumn 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.

Analyzes that fall into the Generalized Linear Models framework at Logistic and Probit regression ($y|X$ has a Binomial distribution), Poisson regression ($y|X$ has a Poisson distribution, Log-linear models (contingency tables).

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

In all of these examples, $E[Y|X] = \mu$ depends on a linear function of X, e.g. $X\beta$.

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

- 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. It also works well in many situations

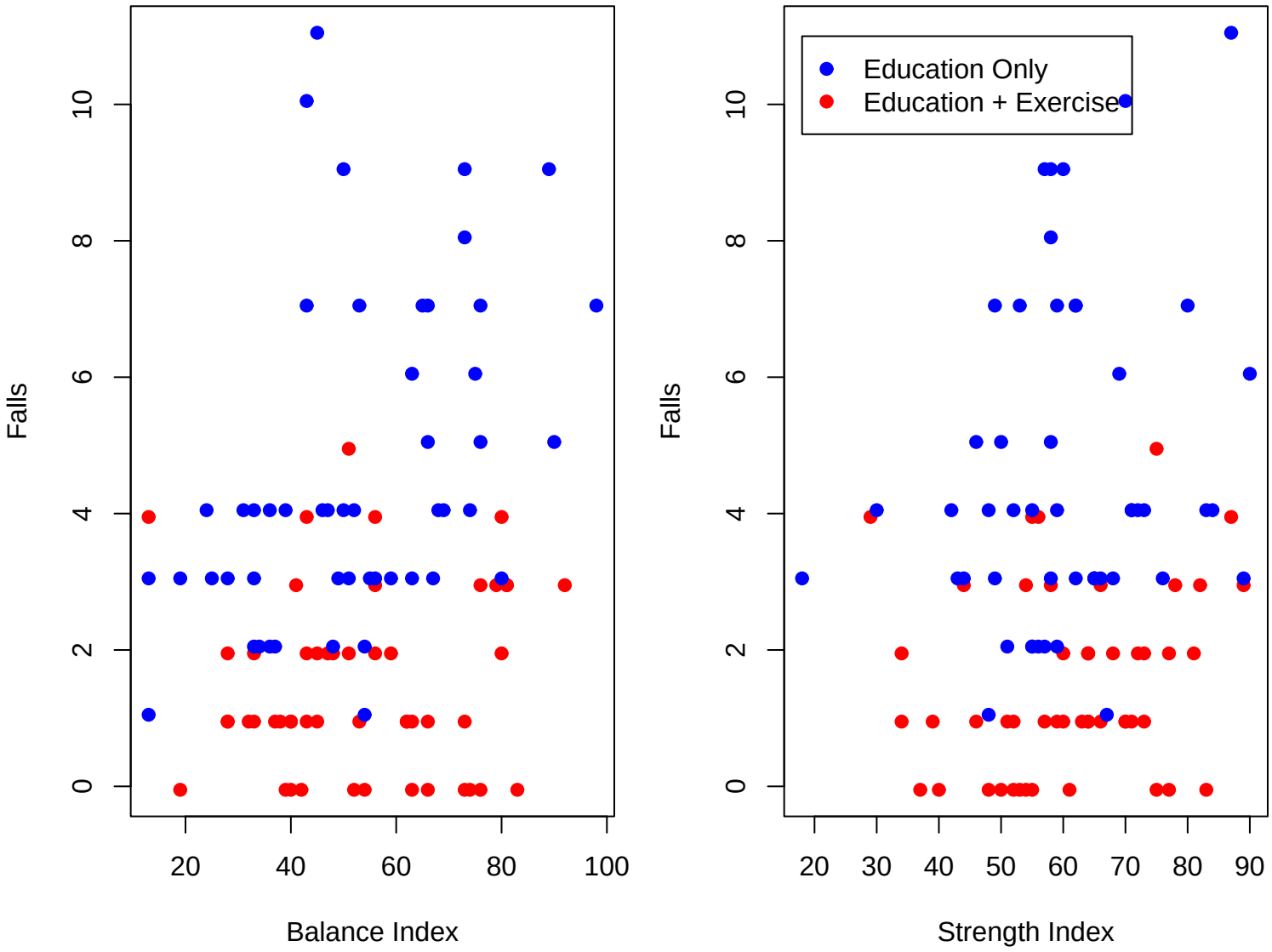
  As mentioned earlier the dispersion parameter $\phi = 1$ is fixed as $\mathrm{Var}(y) = \mu$ for the Poisson distribution.

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)

The question of interest is how much does the treatment reduce the number of falls, after adjusting for Gender, Balance Index and Strength Index


Falls
0
2
4
6
8
10
20
40
60
80
100
Balance Index
Falls
0
2
4
6
8
10
20
30
40
50
60
70
80
90
Strength Index
Education Only
Education + Exercise

- 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 $(-\infty, \infty)$.

For example, the identity link $(g(\mu) = \mu)$ can be used in Poisson regression and in binomial regression problems.

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.

Maximum Likelihood Estimation

In the analysis of Generalized Linear Models, the usual approach to determining parameter estimates is Maximum Likelihood. This is in contrast to linear regression where the usual criterion for optimization is least squares. Though these are the same if you are willing to assume that the $y_i|X_i$ are normally distributed.

Aside:

Often in the regression situation the only assumptions that are made are

- $E[Y_i|X_i] = X_i^T\beta$
- $\mathrm{Var}(Y_i|X_i) = \sigma^2$

There are often no assumptions about the exact distribution of the y's. Without this, maximum likelihood approaches are not possible.

MLE Framework:

Lets assume that that $y_1, y_2, \ldots, y_n$ are conditionally independent given $X_1, X_2, \ldots, X_n$ and a parameter θ (which may be a vector). Let y_i have density (if continous) or mass function (if discrete) $f(y_i|\theta)$.

Then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

The maximum likelihood estimate (MLE) of θ is

$$\hat{\theta} = \arg\sup L(\theta)$$

i.e. the value of θ that maximizes the likelihood function. One way of thinking of the MLE is that its the value of the parameter that is most consistent with the data.

Note that when determining MLEs, it is usually easier to work with the log likelihood function

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(y_i|\theta)$$

It has the same optimum since $\log$ is an increasing function and it is easier to work with since derivatives of sums are usually much nicer than derivatives of products.

Example: Normal mean and variance

Assume $y_1, y_2, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then

$$l(\mu, \sigma) = \sum_{i=1}^{n} \left(-\log 2\pi - \log \sigma - \frac{(y_i - \mu)^2}{2\sigma^2} \right)$$

$$= c - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2$$

To find the MLE of μ and σ, differentiating and setting to 0 is usually the easiest approach. This gives

$$\frac{\partial l(\mu, \sigma)}{\partial \mu} = \frac{-1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 = \frac{-n(\bar{y} - \mu)}{\sigma^2}$$

$$\frac{\partial l(\mu, \sigma)}{\partial \sigma} = \frac{-n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \mu)^2$$

These imply

$$\hat{\mu} = \bar{y}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{n-1}{n}s^2$$

MLEs have nice properties, assuming some regularity conditions. These include

- Consistent (asymptotically unbiased)
- Minimum variance
- Asymptotically normal