Assumptions of linear regression

GLM overview

- In statistics, a generalized linear model (GLM) is a flexible generalization of ordinary linear regression. The GLM generalizes linear regression by allowing the predictor be related to the response variable via a **link function**
- Generalized linear models were formulated as a way of unifying various other statistical models, including linear regression, logistic regression and Poisson regression.
- GLMs use **maximum likelihood estimation** in contrast **to ordinary linear regression** which uses least square for parameter estimation.
- Ordinary linear regression predicts the expected value of a given unknown quantity (**the response variable**, a random variable) as a linear combination of a set of observed values (**predictors**).
- This **implies that a constant change in a predictor leads to a constant change in the response variable** (i.e. a linear-response model).
- This is appropriate when the response variable can vary, to a good approximation, indefinitely in either direction, or more generally for any quantity that only varies by a relatively small amount compared to the variation in the predictive variables, e.g. human heights.
- However, these assumptions are inappropriate for some types of response variables. Ex the beach example and the probability example
- Generalized linear models cover all these situations by allowing for response variables that have arbitrary distributions (rather than simply normal distributions), and for an arbitrary function of the response variable (the link function) to vary linearly with the predictors (rather than assuming that the response itself must vary linearly).
- For example, the case above of predicted number of beach attendees would typically be modeled with a Poisson distribution and a log link, while the case of predicted probability of beach attendance would typically be modeled with a Bernoulli distribution (or binomial distribution, depending on exactly how the problem is phrased) and a log-odds (or logit) link function.
- AAA transformed to a linear relation
- Using the link function

The GLM consists of three elements:

1. An exponential family of probability distributions.

2. A linear predictor

3. A link function

The linear predictor is the quantity which incorporates the information about the independent variables into the model. The symbol η (Greek "eta") denotes a linear predictor. It is related to the expected value of the data through the link function.

η is expressed as linear combinations (thus, "linear") of unknown parameters β.

Link function

The link function provides the relationship between the linear predictor and the mean of the distribution function. There are many commonly used link functions, and their choice is informed by several considerations. There is always a well-defined canonical link function which is derived from the exponential of the response's density function. However, in some cases it makes sense to try to match the domain of the link function to the range of the distribution function's mean, or use a non-canonical link function for algorithmic purposes, for example Bayesian probit regression.

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential Gamma	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0, K)$ K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

Overview [edit]

In a generalized linear model (GLM), each outcome **Y** of the dependent variables is assumed to be generated from a particular distribution in an exponential family, a large class of probability distributions that includes the normal, binomial, Poisson and gamma distributions, among others. The mean, $\boldsymbol{\mu}$, of the distribution depends on the independent variables, **X**, through:

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$$

where E(**Y**|**X**) is the expected value of **Y** conditional on **X**; $\mathbf{X}\boldsymbol{\beta}$ is the *linear predictor*, a linear combination of unknown parameters $\boldsymbol{\beta}$; g is the link function.

In this framework, the variance is typically a function, **V**, of the mean:

$$\mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \mathrm{V}(g^{-1}(\mathbf{X}\boldsymbol{\beta})).$$

It is convenient if **V** follows from an exponential family of distributions, but it may simply be that the variance is a function of the predicted value.

The unknown parameters, $\boldsymbol{\beta}$, are typically estimated with maximum likelihood, maximum quasi-likelihood, or Bayesian techniques.

Fitting

Maximum likelihood

The maximum likelihood estimates can be found using an iteratively reweighted least squares algorithm

Bayesian methods

In general, the posterior distribution cannot be found in closed form and so must be approximated, usually using Laplace approximations or some type of Markov chain Monte Carlo method such as Gibbs sampling.

NOT least square (OLS)

Linear regression

- A simple, very important example of a generalized linear model (also an example of a general linear model) **is linear regression.** In linear regression, the use of the least-squares estimator is justified by the Gauss–Markov theorem, which does not assume that the distribution is normal.
- the link function is the **identity**

Binary data

When the response data, Y, are binary (taking on only values 0 and 1), the distribution function is generally chosen to be the Bernoulli distribution and the interpretation of μi is then the probability, p, of Yi taking on the value one.

There are several popular link functions for binomial functions.

Logit link function

The most typical link function is the canonical logit link:

GLMs with this setup are logistic regression models (or logit models).

Identity link

The identity link g(p) = p is also sometimes used for binomial data to yield a linear probability model. However, the identity link can predict nonsense "probabilities" less than zero or greater than one. This can be avoided by using a transformation like cloglog, probit or logit (or any inverse cumulative distribution function). A primary merit of the identity link is that it can be estimated using linear math—and other standard link functions are approximately linear matching the identity link near p = 0.5.

Multinomial regression

The binomial case may be easily extended to allow for a multinomial distribution as the response (also, a Generalized Linear Model for counts, with a constrained total). There are two ways in which this is not usually done:

Ordered response

If the response variable is ordinal, then one may fit a model function of the form:

Different links g lead to ordinal regression models like proportional odds models or ordered probit models.

Count data

Another example of generalized linear models includes Poisson regression which models count data using the Poisson distribution. The link is typically the logarithm, the canonical link.