

Introduction to Generalized Linear Models

UCR GradQuant Workshop
5/28/2015

Roadmap



- General linear model
- Generalized linear model
- Logistic regression model
 - An example using R

General Linear Model

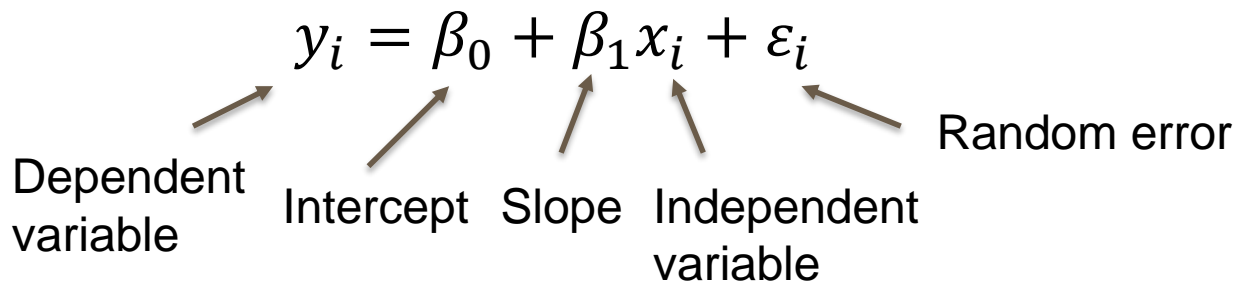


In a general linear model, the **response** y_i , $i = 1, \dots, n$ is modelled by a linear function of **independent** variables x_j , $j = 1, \dots, p$ plus an error term.

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \varepsilon_i$$

General Linear Model

Here general refers to the dependence on potentially more than one explanatory variable, vs. the simple linear model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$


Dependent variable Intercept Slope Independent variable Random error

The model is linear in the parameters, e.g.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

but not e.g.

$$y_i = \beta_0 + \beta_1 x_i^{\beta_2} + \varepsilon_i$$

General Linear Model



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

$$\begin{aligned}E[\varepsilon_i] &= 0 \\ \text{var}[\varepsilon_i] &= \sigma^2\end{aligned}$$

Typically we assume

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

as a basis for inference, e.g. t-tests on parameters. The errors are uncorrelated.

General Linear Model



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)
- › 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 Model



Comparison

	General Linear Model	Generalized Linear Model
Special cases	ANOVA, ANCOVA, MANOVA, MANCOVA, linear regression, mixed model	Linear regression, logistic regression, Poisson regression
Function in R	Lm()	Glm()
Typical method estimation	Least Square	Maximum Likelihood

Generalized Linear Model



A generalized linear model is made up of a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

and two functions

- › a link function that describes how the mean, $E[Y_i] = \mu_i$ depends on the linear predictor

$$\eta_i = g[\mu_i]$$

- › a variance function that describes how the variance, $var[Y_i]$, depends on the mean

$$var[Y_i] = \phi V(\mu_i)$$

where the dispersion parameter ϕ is a constant

Generalized Linear Model



› Normal General Linear Model as a Special Case

For the general linear model with $\varepsilon \sim N(0, \sigma^2)$ we have the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

the link function

$$g[\mu_i] = \mu_i$$

and the variance function

$$V[\mu_i] = 1$$

Generalized Linear Model



› Modelling Binomial Data

Suppose

$$Y_i \sim \text{Binomial}(n_i, p_i)$$

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

$$E[Y_i/n_i] = p_i \quad \text{var}[Y_i/n_i] = p_i(1 - p_i)/n_i$$

So our variance function is

$$V[\mu_i] = \mu_i(1 - \mu_i)$$

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

$$g[\mu_i] = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right)$$

Generalized Linear Model



› Modelling Poisson Data

Suppose

$$Y_i \sim \text{Poisson}(\lambda_i)$$

Then

$$E[Y_i] = \lambda_i \quad \text{var}[Y_i] = \lambda_i$$

So our variance function is

$$V[\mu_i] = \mu_i$$

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

$$g[\mu_i] = \log(\mu_i)$$

Binary Data



Binary data may occur in two forms

- › ungrouped in which the variable can take one of two values, say success/failure
- › grouped in which the variable is the number of successes in a given number of trials

The natural distribution for such data is the *Binomial*(n, p) distribution, where in the first case $n = 1$

Models for Binary Data

We saw previously that Binomial data may be modelled by a generalized linear model with logit link. This model is known as the logistic regression model and is the most popular for binary data.

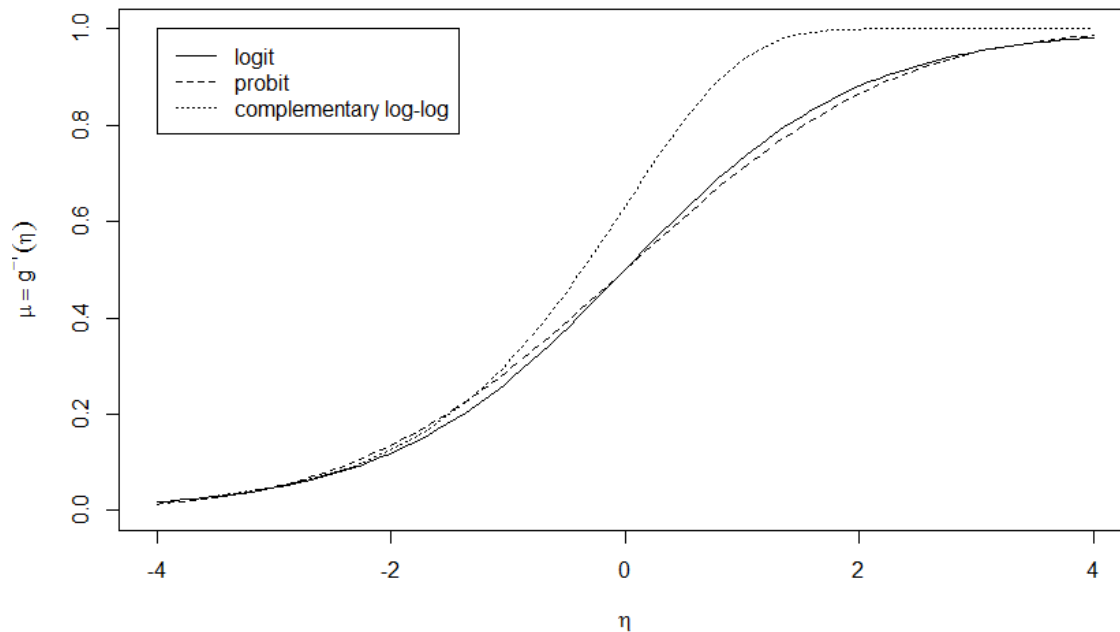
There are two other links commonly used in practice:

- ▶ probit link $g[\mu_i] = \Phi^{-1}(\mu_i)$ where Φ denotes the cumulative distribution function of $N(0, 1)$
- ▶ complementary log-log link $g[\mu_i] = \log(-\log(1 - \mu_i))$

Binary Data

Comparison of Links

The three links map the linear predictor to the probability scale as follows:



Binary Data



Choice of Link

The logit and probit functions are symmetric and - once their variances are equated - are very similar. Therefore it is usually difficult to choose between them on the grounds of fit.

The logit is usually preferred over the probit because of its simple interpretation as the logarithm of the odds of success $p_i/(1 - p_i)$.

The complementary log-log is asymmetric and may therefore be useful when the logit and probit links are inappropriate.

We will concentrate on using the logit link.

Logit Transformation

- First, we move from the probability p_i to the *odds*

$$odds_i = \frac{p_i}{1 - p_i}$$

defined as the ratio of the probability to its complement, or the ratio of favorable to unfavorable cases.

- If the probability of an event is a half, the odds are one-to-one or even
- Odd can take any positive value.

Logit Transformation

- › Second, we take logarithms, calculating the *logit* or log-odds

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

which has the effect of removing the floor restriction.

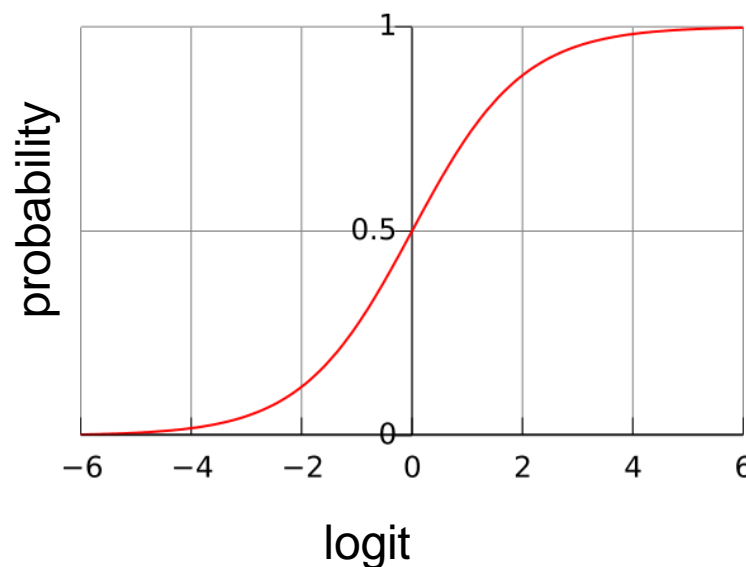
- › If the probability of an event is a half, the odds are even and logit is zero.
- › Negative logits represent probabilities below on half and positive logits correspond to probabilities above on half.

Logistic Regression

› Logistic function

$$\eta_i = \log\left(\frac{p_i}{1-p_i}\right) \Rightarrow p_i = f(\eta_i) = \frac{1}{1+e^{-\eta_i}}$$

- › It can take as an input any value from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1.
- › Variable η could represent the exposure to some set of risk factors
- › $f(\eta_i)$ represents the probability of a particular outcome, given that set of risk factors.



Logistic Regression

› Logistic function

- › The variable η is a measure of the total contribution of all independent variables used in the model

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- › Each of the regression coefficients describes the size of the contribution of that independent variables
 - › A positive value means that that independent variable increases the probability of the outcome;
 - › A negative value means that independent variable decreases the probability of the outcome.

Logistic Regression

› Fitting the logistic regression model

- › Criteria: Find values of unknown parameters that maximize the probability of obtaining the observed set of data.
- › Likelihood function: express the probability of the observed data as a function of the unknown parameters
 - › The probability of a pair of observation $(x_{1i}, \dots, x_{ki}, y_i)$

$$p(x_i)^{y_i} \cdot (1 - p(x_i))^{1-y_i} \quad P(Y_i = 1) = p(x_i)$$

- › The observations are assumed to be independent, the likelihood function has be expressed as follows:

$$l(\beta) = \prod_{i=1}^n p(x_i)^{y_i} \cdot (1 - p(x_i))^{1-y_i}$$

Logistic Regression

› Fitting the logistic regression model

› Maximum Likelihood (ML)

› *Log likelihood*

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^n [y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i))]$$

- › Maximize log likelihood function by differentiating $L(\beta)$ with respect to β and setting the resulting expressions equal to zero

$$\sum_i (y_i - p(x_i)) = 0$$

$$\sum_i x_{k,i} (y_i - p(x_i)) = 0$$

- › The value of β given by the solution to the above equations is called the maximum likelihood estimate and will be denoted as $\hat{\beta}$

Deviance

Test the goodness of fit

- › Deviance: measures how close the predicted values from the fitted model match the actual values from the raw data.

$$D = -2[\log\text{-likelihood}(\text{proposed model}) - \log\text{-likelihood}(\text{saturated model})]$$

- › A saturated model is a model that fits the data perfectly, so its log-likelihood is the maximum. It has as many parameters as observations and hence it provides no simplification at all.
- › The deviance has a chi-squared asymptotic null distribution.
- › The degree of freedom is $n-p$, where n is the number of observations and p is the number of model parameters.
- › Smaller deviance, the better the model.

Wald Test

- › Testing the significance of the coefficients
 - › Wald Test: a Wald test calculates a Z statistic, which is

$$z = \frac{\hat{\beta}}{\hat{se}(\hat{\beta})}$$

under the hypothesis that $\beta = 0$, the resulting ratio will follow a standard normal distribution. This z value is then squared, yielding a Wald statistic with a chi-square distribution.

Interpretation of Logistic Models

Consider the logistic model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

If we increase x by one unit

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 (x_i + 1) = \beta_0 + \beta_1 x_i + \beta_1$$

$$\left(\frac{p_i}{1-p_i}\right) = \exp(\beta_0 + \beta_1 x_i) \exp(\beta_1)$$

the odds are multiplied by $\exp(\beta_1)$.

Example: Budworm Data

Collett(1991) describes an experiment on the toxicity of the pyrethoid *trans* - *cypermethrin* to the tobacco budworm. Batches of 20 moths of each sex were exposed to varying doses of the pyrethoid for three days and the number knocked out in each batch was recorded:

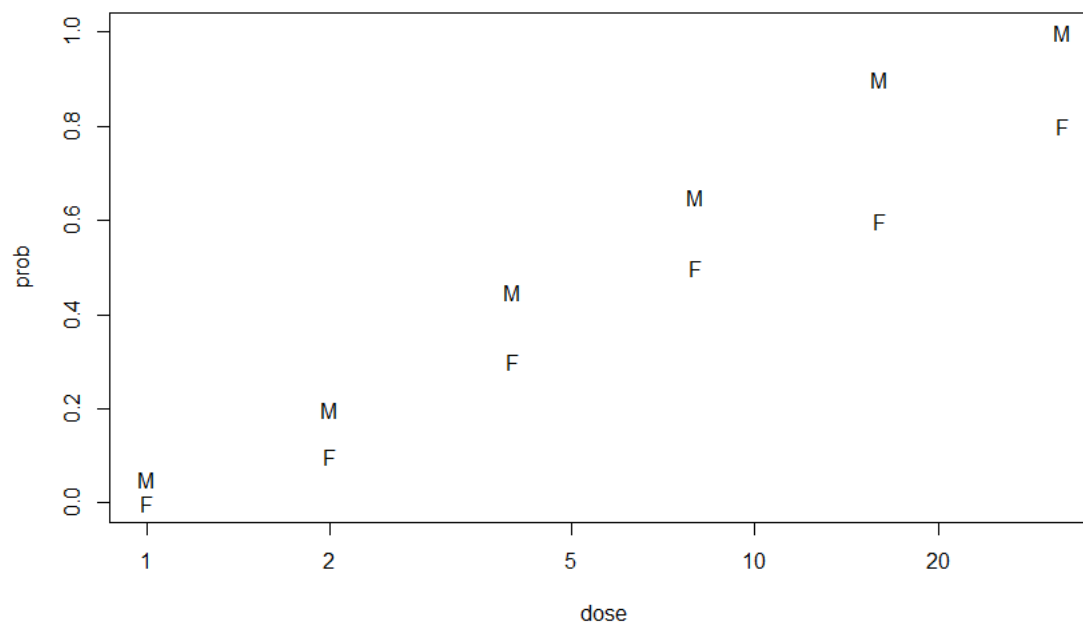
sex	Dose (μg)					
	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

Since the doses are in powers of two, we will use $\log_2(\text{dose})$ as the response.

Example: Budworm Data

Scatterplots of Binomial Data

For grouped binary data, scatterplots are more helpful:

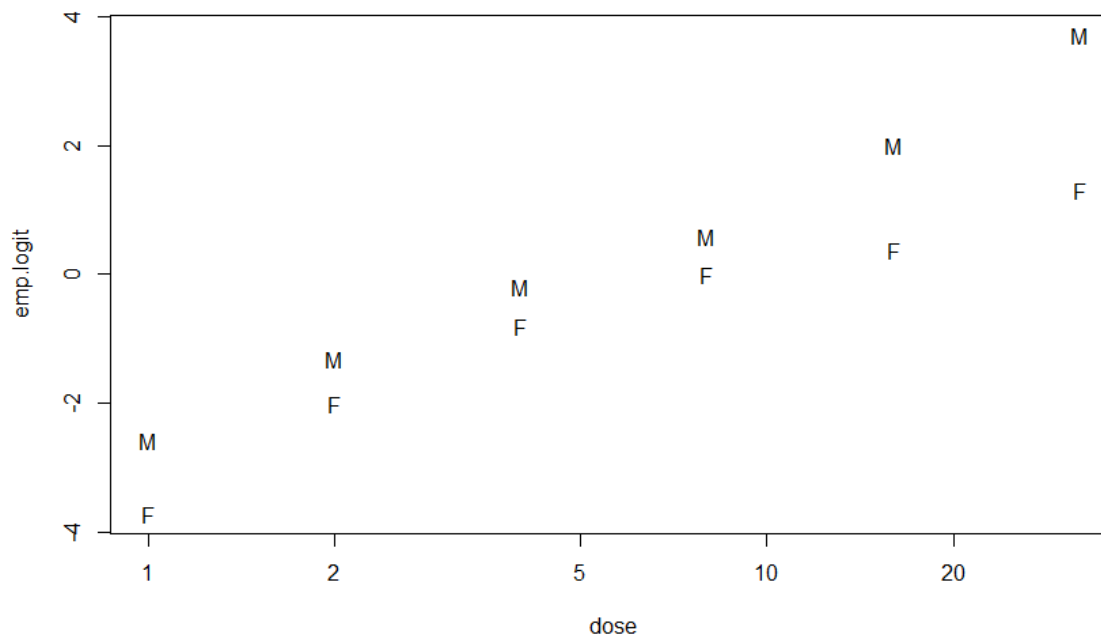


Example: Budworm Data

Scatterplot Scales

When fitting a logistic model, it can also be helpful to plot the data on the logit scale. To avoid dividing by zero, we calculate the empirical logits

$$\log\left(\frac{(y_i + 0.5)/n_i}{1 - (y_i + 0.5)/n_i}\right) = \log\left(\frac{y_i + 0.5}{n_i + 0.5 - y_i}\right)$$



Binomial Responses and glm

Now we would like to fit our candidate models. Binomial responses can be specified to glm in three ways:

- › a numeric vector giving the proportion of successes y_i/n_i , in which case a vector of the prior weights n_i must be passed to the weights argument
- › a numeric 0/1 vector (0 = failure); a logical vector (FALSE = failure), or a factor (first level = failure)
- › a two-column matrix with the number of successes and the number of failures

Better starting values are generated when the third format is used.

Example: Budworm Data

Modelling the data

A linear logistic model appears to be appropriate. A reasonable approach might be to consider the following linear predictors:

- › single line for both sexes ($\sim \text{ldose}$)
- › parallel lines for each sex ($\sim \text{ldose} + \text{sex}$)
- › separate lines for each sex ($\sim \text{ldose} + \text{sex} + \text{ldose}:\text{sex}$)

How can we determine which model is best?

Nested Models

- › The candidate models for the budworm data are an example of nested models where each model is a special case of the models that have a greater number of terms.
- › We can compare nested models by testing the hypothesis that some of the parameters of a larger model are equal to zero.

Nested Models

For example suppose we have the model

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

we can test

$$H_0: \beta_{q+1} = \cdots = \beta_p = 0$$

versus $H_1: \beta_j \neq 0$, for some $j \in \{q + 1, p\}$

using the likelihood ratio statistic

$$LR = 2(l_{big} - l_{small})$$

where l_m is the maximized log-likelihood under model m , i.e. $l(\widehat{\beta}_m)$. Under the null hypothesis, LR is approximately χ_d^2 where $d = p - q$.

Example: Budworm Data

Modelling the data

- › Single line model
 - › As expected, the coefficient of Idose is highly significant.
- › Parallel lines model
 - › the Wald tests suggest both Idose and sex are needed in the model.
 - › Likelihood ratio test can also be used to compare models and we can use anova to perform this test.
- › separate lines model
 - › Using anova will test sequential addition of terms in this model.
 - › Allowing separate slopes does not significantly reduce deviance.

Goodness-of-fit

- › The parallel lines model has a deviance of 6.76 on 9 degrees of freedom, indicating that the model fits well.

Example: Budworm Data

For the budworm data, the parallel lines model is

$$\log\left(\frac{p_i}{1-p_i}\right) = -3.47 + 1.06 \text{dose}_i + 1.10(\text{sex}_i == \text{"M"})$$

Therefore

- › the odds of death for a male moth are $\exp(1.10) = 3.01$ times that for a female moth, given a fixed dose of the pyrethroid.
- › the odds of death increase by a factor of $\exp(1.06) = 2.90$ for every $\log \mu g$ of pyrethroid, for male or female moths.

Example: Budworm Data

Wald Confidence Intervals

Confidence intervals for the parameters can be based on the asymptotic normal distribution for $\hat{\beta}_j$.

For example a 95% confidence interval would be given by

$$\hat{\beta}_j \pm 1.96 * s.e.(\hat{\beta}_j)$$

Such confidence intervals can be obtained as follows:

```
confint.lm(parr)
```

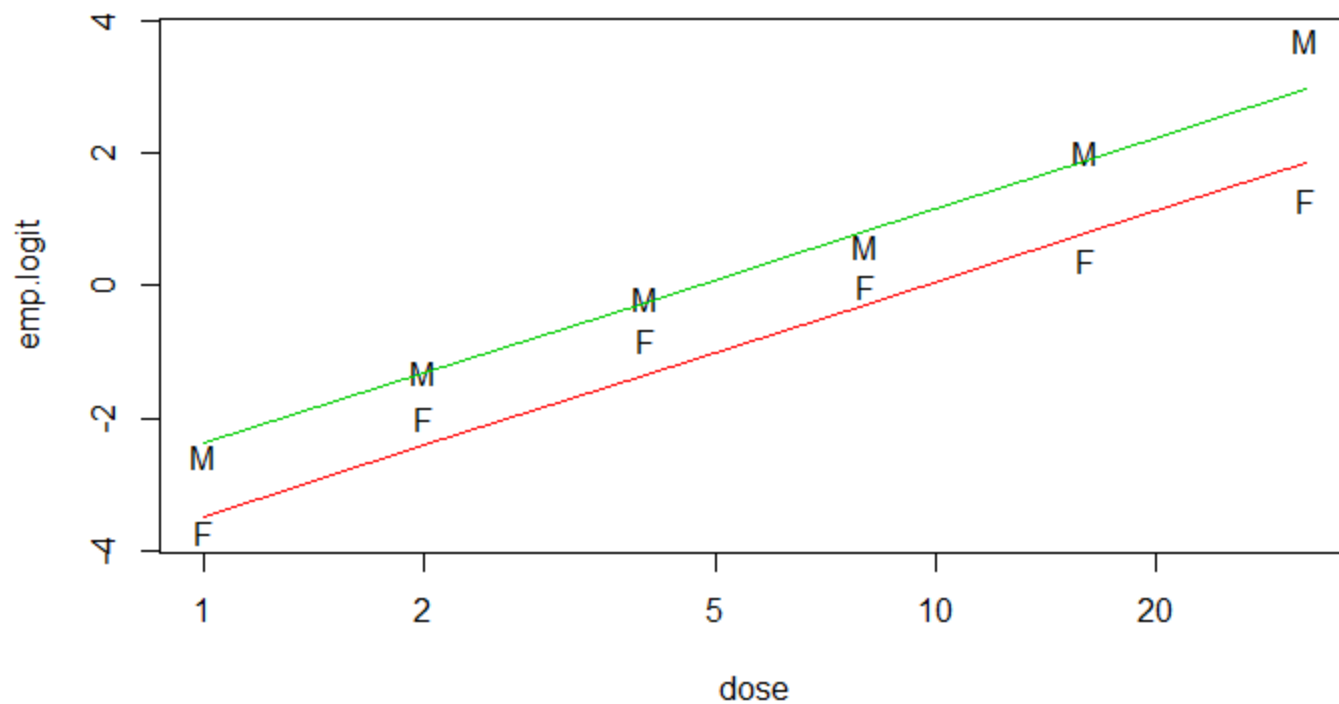
Prediction

The predict method for GLMs has a type argument, which may be specified as

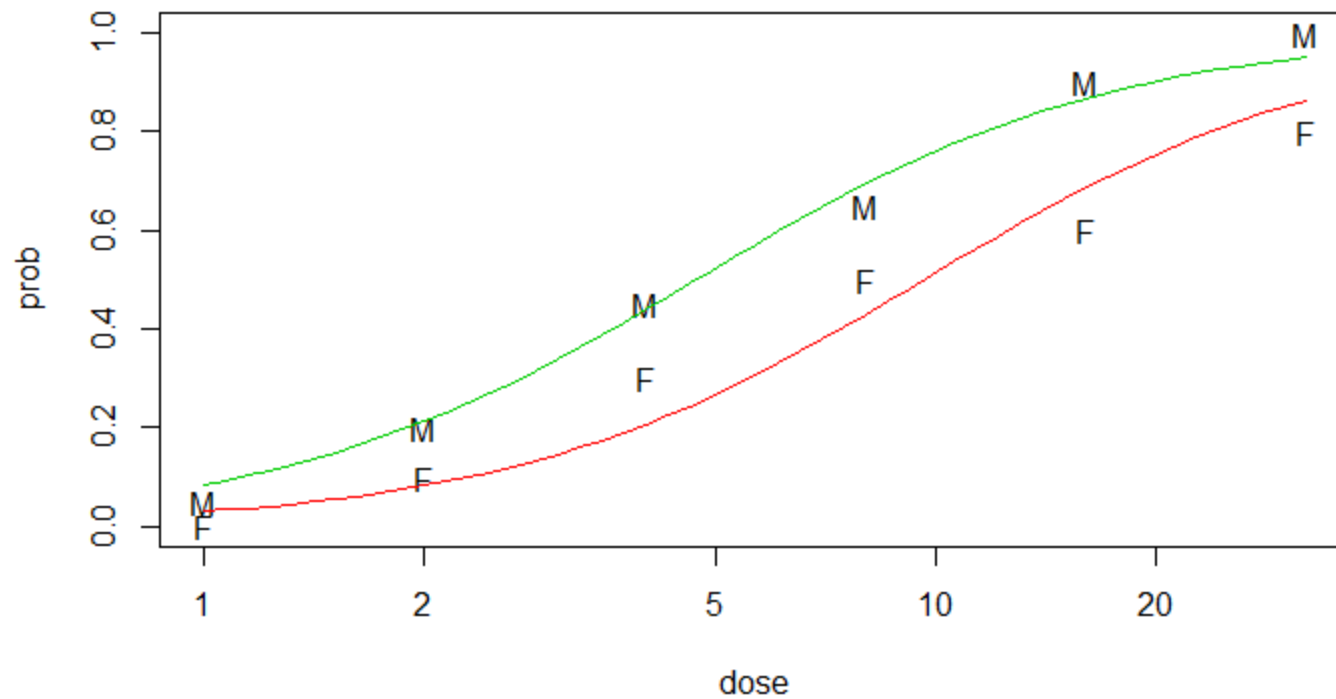
- › "link" for predictions of η
- › "response" for predictions of μ

If no new data is passed to predict, these options return `object$linear.predictor` and `object$fitted.values` respectively.

Example: Budworm Data



Example: Budworm Data



Thanks!