The R Statistical Computing Environment
Basics and Beyond
Linear and Generalized Linear Models in R

John Fox

McMaster University

ICPSR/Berkeley 2016

Linear and Generalized Linear Models in R

Topics

To be covered as time permits:

- Multiple linear regression
- Factors and dummy regression models
- Overview of the `lm` function
- The structure of generalized linear models (GLMs) in R; the `glm` function
- GLMs for binary/binomial data
- GLMs for count data

Linear Models in R

Arguments of the lm function

- `lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)`
- `formula`

Expression	*Interpretation*	*Example*
`A + B`	include both `A` and `B`	`income + education`
`A - B`	exclude `B` from `A`	`a*b*d - a:b:d`
`A:B`	all interactions of `A` and `B`	`type:education`
`A*B`	`A + B + A:B`	`type*education`
`B %in% A`	`B` nested within `A`	`education %in% type`
`A/B`	`A + B %in% A`	`type/education`
`A^k`	effects crossed to order `k`	`(a + b + d)^2`

Linear Models in R

Arguments of the lm function

- `data`: A data frame containing the data for the model.
- `subset`:
  - a logical vector: `subset = sex == "F"`
  - a numeric vector of observation indices: `subset = 1:100`
  - a negative numeric vector with observations to be omitted: `subset = -c(6, 16)`
- `weights`: for weighted-least-squares regression
- `na.action`: name of a function to handle missing data; default given by the `na.action` option, initially `"na.omit"`
- `method`, `model`, `x`, `y`, `qr`, `singular.ok`: technical arguments
- `contrasts`: specify list of contrasts for factors; e.g., `contrasts=list(partner.status=contr.sum, fcategory=contr.poly))`
- `offset`: term added to the right-hand-side of the model with a fixed coefficient of 1.

Generalized Linear Models in R

Review of the Structure of GLMs

- A generalized linear model consists of three components:

1. A *random component*, specifying the conditional distribution of the response variable, y_i, given the predictors. Traditionally, the random component is an exponential family — the normal (Gaussian), binomial, Poisson, gamma, or inverse-Gaussian.
2. A linear function of the regressors, called the *linear predictor*,

$$\eta_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

   on which the expected value μ_i of y_i depends.
3. A *link function* $g(\mu_i) = \eta_i$, which transforms the expectation of the response to the linear predictor. The inverse of the link function is called the *mean function*: $g^{-1}(\eta_i) = \mu_i$.

Generalized Linear Models in R

Review of the Structure of GLMs

- In the following table, the logit, probit and complementary log-log links are for binomial or binary data:

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
identity	μ_i	η_i
log	$\log_e \mu_i$	e^{η_i}
inverse	μ_i^{-1}	η_i^{-1}
inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
square-root	$\sqrt{\mu_i}$	η_i^2
logit	$\log_e \dfrac{\mu_i}{1-\mu_i}$	$\dfrac{1}{1+e^{-\eta_i}}$
probit	$\Phi(\mu_i)$	$\Phi^{-1}(\eta_i)$
complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

Generalized Linear Models in R

Implementation of GLMs in R

- Generalized linear models are fit with the `glm` function. Most of the arguments of `glm` are similar to those of `lm`:
  - The response variable and regressors are given in a model `formula`.
  - `data`, `subset`, and `na.action` arguments determine the data on which the model is fit.
  - The additional `family` argument is used to specify a *family-generator function*, which may take other arguments, such as a link function.

Generalized Linear Models in R

Implementation of GLMs in R

- The following table gives family generators and default links:

Family	*Default Link*	*Range of* y_i	$V(y_i \mid \eta_i)$
`gaussian`	`identity`	$(-\infty, +\infty)$	ϕ
`binomial`	`logit`	$\dfrac{0, 1, \ldots, n_i}{n_i}$	$\mu_i(1-\mu_i)$
`poisson`	`log`	$0, 1, 2, \ldots$	μ_i
`Gamma`	`inverse`	$(0, \infty)$	$\phi\mu_i^2$
`inverse.gaussian`	`1/mu^2`	$(0, \infty)$	$\phi\mu_i^3$

- For distributions in the exponential families, the variance is a function of the mean and a dispersion parameter ϕ (fixed to 1 for the binomial and Poisson distributions).

Generalized Linear Models in R
Implementation of GLMs in R

- The following table shows the links available for each family in R, with the default links as ■:

family	link			
	identity	inverse	sqrt	1/mu^2
gaussian	■	□		
binomial				
poisson	□		□	
Gamma	□	■		
inverse.gaussian	□	□		■
quasi	■	□	□	□
quasibinomial				
quasipoisson	□		□	

Generalized Linear Models in R
Implementation of GLMs in R

family	link			
	log	logit	probit	cloglog
gaussian	□			
binomial	□	■	□	□
poisson	□			
Gamma	□			
inverse.gaussian	□			
quasi	□	□	□	□
quasibinomial		■	□	□
quasipoisson	■			

- The `quasi`, `quasibinomial`, and `quasipoisson` family generators do not correspond to exponential families.

Generalized Linear Models in R
GLMs for Binary/Binomial and Count Data

- The response for a binomial GLM may be specified in several forms:
  - For binary data, the response may be
    - a variable or an S expression that evaluates to `0`'s ('failure') and `1`'s ('success').
    - a logical variable or expression (with `TRUE` representing success, and `FALSE` failure).
    - a factor (in which case the first category is taken to represent failure and the others success).
  - For binomial data, the response may be
    - a two-column matrix, with the first column giving the count of successes and the second the count of failures for each binomial observation.
    - a vector giving the *proportion* of successes, while the binomial denominators (total counts or numbers of trials) are given by the `weights` argument to `glm`.

Generalized Linear Models in R
GLMs for Binary/Binomial and Count Data

- Poisson generalized linear models are commonly used when the response variable is a count (Poisson regression) and for modeling associations in contingency tables (loglinear models).
  - The two applications are formally equivalent. Poisson GLMs are fit in S using the `poisson` family generator with `glm`.
- Overdispersed binomial and Poisson models may be fit via the `quasibinomial` and `quasipoisson` families.