

R Formulas Notes

Matthew Turner

Department of Psychology

Georgia State University

Formulas

- R has a shorthand formula language
 - Developed in the 1980's by John Chambers
 - It is designed to make it easy to enter statistical models
 - It is used for all linear models (ANOVA, Regression) and simple extensions of the language cover mixed and hierarchical models (**lme4**, **nlme** packages)

Basic Idea

- A model is specified as:

Dependent_variable ~ Independent Variable(s)

- You just list the variables as they appear in your mathematical notation
- The constant term (in regression) is assumed

Basic Idea

- Example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\mathbf{y} \sim \mathbf{x1} + \mathbf{x2}$$

- If you wanted no intercept:

$$y = \beta_1 x_1 + \beta_2 x_2$$

$$\mathbf{y} \sim \mathbf{-1} + \mathbf{x1} + \mathbf{x2}$$

Symbol	Example	Meaning
+	+X	include this variable
-	-X	delete this variable
:	X:Z	include the interaction between these variables
*	X*Y	include these variables and the interactions between them
	X Z	conditioning: include x given z
^	(X + Z + W) ^ 3	include these variables and all interactions up to three way
I	I (X*Z)	as is: include a new variable consisting of these variables multiplied
1	X - 1	intercept: delete the intercept (regress through the origin)

$Y \sim X + Z + W + X:Z + X:W + Z:W$

$Y \sim X * Z * W - X:Z:W$

$Y \sim (X + Z + W) ^ 2$

Symbol	Example	Meaning
+	+X	include this variable
-	-X	delete this variable
:	X:Z	include the interaction between these variables
*	X*Y	include these variables and the interactions between them
	X Z	conditioning: include x given z
^	(X + Z + W) ^ 3	include these variables and all interactions up to three way
I	I (X*Z)	as is: include a new variable consisting of these variables multiplied
1	X - 1	intercept: delete the intercept (regress through the origin)

$$Y \sim X + Z + W + X:Z + X:W + Z:W$$

$$Y \sim X * Z * W - X:Z:W$$

$$Y \sim (X + Z + W) ^ 2$$

Variable Types Determine Models

- For the model: $y \sim x1 + x2$
- If $x1$ and $x2$ are categorical then it is an ANOVA
- If $x1$ and $x2$ are numerical then it is a regression
- If $x1$ is categorical and $x2$ is numerical then it is an ANCOVA

Resources

- The following are good summaries of the model formulae – look at all of them and pick the one(s) that you like best:
 - <https://ww2.coastal.edu/kingw/statistics/R-tutorials/formulae.html> ([costal.edu](https://ww2.coastal.edu/kingw/statistics/) has many other introductory articles, too!)
 - <https://science.nature.nps.gov/im/datamgmt/statistics/r/formulas/>
 - <http://conjugateprior.org/2013/01/formulae-in-r-anova/> (this page has many examples of ANOVA and mixed-models)
 - More advanced:
http://genomicsclass.github.io/book/pages/expressing_design_formula.html this covers the relationship from formula to the design matrix for linear models