Big Data: Data Analysis Boot Camp
Formulae Notation

Chuck Cartledge, PhD

19 January 2018

Table of contents (1 of 1)

What are we going to cover?

1. Look at the basic ideas behind R's formula object
2. Look at how sample R formulae can be represented using traditional mathematical notation
3. Use an attached program to create linear regression error terms and how they can be displayed

Basic idea

"When discussing models, the term linear does not mean a straight-line. Instead, a linear model contains additive terms, each containing a single multiplicative parameter; thus, the equations

$$\begin{aligned} y &= \beta_0 + \beta_1 x \\ y &= \beta_0 + \beta_1 x^1 + \beta_2 x^2 \\ y &= \beta_0 + \beta_1 x^2 \\ y &= \beta_0 + \beta_1 x^1 + \beta_2 \log(x^2) \end{aligned}$$

are linear models. The equation $y = \alpha x^\beta$, however, is not a linear model."[1]

[1] Author unknown, see attached file.

A little "magic" to confuse things

"The definition of a linear model is an equation that contains mathematical variables, parameters and random variables and that is linear in the parameters and in the random variables. What this means is that if a, b and c are parameters then obviously

$$y = a + bx$$

is a linear model, but so is

$$y = a + bx - cx^2$$

because x^2 can be replaced by z which gives a linear relationship

$$y = a + bx - cz$$

and so is

$$y = a + b\exp(x)$$

because we can create a new variable $z = \exp(x)$, so that

$$y = a + bz$$

Some models are non-linear but can be readily linearized by transformation. For example:

$$y = \exp(a + bx)$$

is non-linear, but on taking logs of both sides, it becomes

$$\ln(y) = a + bx"$$

M. Crawley [1]

Basic operators and ideas

R uses the idea of a formula object to guide and direct modeling scripts. Formula notation and operators have different meanings than mathematical operators.[1]

response variable ˜explanatory variable(s)

Operator	Meaning
˜	"is modeled as a function of"
+	separate explanatory terms (not addition)
:	separate variable and factor names
*	indicates inclusion of explanatory variables and interactions (not multiplication)
ˆ	crossing to the specified degree
%in%	terms on the left are nested in those on the right
-	removes specified terms (not subtraction)
func	mathematical functions can be used on response or explanatory variables
I()	identify portions of formula to be used in their mathematical sense
.	use all columns not otherwise in the formula
/	indicates nesting of explanatory variables in the model
|	indicates conditioning (not 'or'), so that $y \tilde{} x \mid z$ is read as "y as a function of x given z"

Lots of examples (1 of 5)[1]

Model	**Syntax**	**Math.**	**Comments.**
Null	y ˜1	–	1 is the intercept in regression models, but here it is the overall mean y
Regression	y ˜x	$y = \beta_0 + \beta_1 x$	x is a continuous explanatory variable
Regression through origin	y ˜$x - 1$	$y = \beta_1 x$	Do not fit an intercept
One-way ANOVA	y ˜*sex*	–	sex is a two-level categorical variable
One-way ANOVA	y ˜*sex* $- 1$	–	as above, but do not fit an intercept (gives two means rather than a mean and a difference)

Lots of examples (2 of 5)[1]

Model	Syntax	Math.	Comments.
Two-way ANOVA	$y\ \tilde{}\ sex + genotype$	–	genotype is a four-level categorical variable
Factorial ANOVA	$y\ \tilde{}\ N * P * K$	–	N, P and K are two-level factors to be fitted along with all their interactions
Three-way ANOVA	$y\ \tilde{}\ N * P * K - N : P : K$	–	As above, but dont fit the three-way interaction
Analysis of covariance	$y\ \tilde{}\ x + sex$	–	A common slope for y against x but with two intercepts, one for each sex
Nested ANOVA	$y\ \tilde{}\ a/b/c$	–	Factor c nested within factor b within factor a

Lots of examples (3 of 5)[1]

Model	Syntax	Math.	Comments.
Split-plot ANOVA	y ~$a * b * c + Error(a/b/c)$	–	A factorial experiment but with three plot sizes and three different error variances, one for each plot size
Multiple regression	y ~$x + z$	–	Two continuous explanatory variables, flat surface fit
Multiple regression	y ~$x * z$	–	Fit an interaction term as well (x + z + x:z)
Multiple regression	y ~x+I$(x^2) + z$+I(z^2)	–	Fit a quadratic term for both x and z

Lots of examples (4 of 5)[1]

Model	Syntax	Math.	Comments.
Multiple regression	y ~ $poly(x, 2) + z$	–	Fit a quadratic polynomial for x and linear z
Multiple regression	y ~ $(x + z + w)^2$	–	Fit three variables plus all their interactions up to two-way
Non-parametric model	y ~ $s(x) + s(z)$	–	y is a function of smoothed x and z in a generalized additive model
Transformed response and explanatory variables	$log(y)$ ~ $\mathrm{I}(1/x) + sqrt(z)$	–	All three variables are transformed in the model

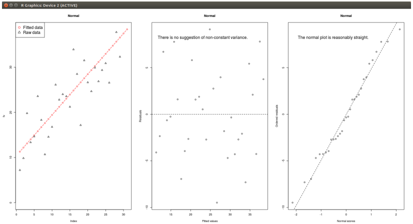
Lots of examples (5 of 5)[1]

Model	Syntax	Math.	Comments.
Polynomial	y ˜x+I(x^2)	$y = \beta_0 + \beta_1 x + \beta_2 x^2$	Polynomial model; note that the identity function I() allows terms in the model to include normal mathematical symbols
First order	y ˜$x + z$	$y = \beta_0 + \beta_1 x + \beta_2 z$	A first-order model in x and z without interaction terms.
First order with interaction	y ˜$x : z$	$y = \beta_0 + \beta_1 xz$	A model containing only first-order interactions between x and z.
First order with term	y ˜$x * z$	$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$	A full first-order model with a term; an equivalent code is y ˜x + z + x:z.
All first order	y ˜$(A + B + C)^2$	$y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 BC$	A model including all first-order effects and interactions up to the nth order, where n is given by $()^n$. An equivalent code in this case is y ˜$A*B*CA : B : C$.

Load and execute the attached script

Load the execute the attached file.
Each plot uses different error types, and shows:

- The raw and fitted data
- Comments about the residual values
- Interpretations about the qqplot



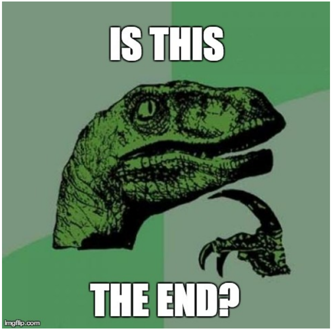
Attached file.

Q & A time.

"A human being should be able to change a diaper, plan an invasion, butcher a hog, conn a ship, design a building, write a sonnet, balance accounts, build a wall, set a bone, comfort the dying, take orders, give orders, cooperate, act alone, solve equations, analyze a new problem, pitch manure, program a computer, cook a tasty meal, fight efficiently, die gallantly. Specialization is for insects."
Robert Heinlein, Time Enough for Love

What have we covered?

- Reviewed R's formula notation and how it differs from traditional mathematical notation
- Looked at how sample R formulae expand into traditional mathematical notation
- Looked at how different error terms can be displayed and detected



References (1 of 1)

[1] Michael J. Crawley, The R Book, John Wiley & Sons, 2012.

Files of interest

1. Modeling different errors
2. Using R for Linear Regression