Statistical Formula Notation in R

R functions, notably `lm()` for fitting linear regressions and `glm()` for fitting logistic regressions, use a convenient formula syntax to specify the form of the statistical model to be fit. The basic format of such a formula is

response variable ~ predictor variables

The tilde is read as “is modeled as a function of.” A basic regression analysis would be formulated as

\[ Y \sim X \]

Therefore we might fit a linear model regressing \( Y \) on \( X \) as

\[ \text{fit} \leftarrow \text{lm}(Y \sim X) \]

where \( X \) is the predictor variable and \( Y \) is the response variable. In the usual mathematical notation this corresponds to the linear regression model denoted

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \]

Additional explanatory variables can be included using the “+” symbol. To add another predictor variable \( Z \), the formula becomes

\[ Y \sim X + Z \]

and the linear regression call becomes

\[ \text{fit} \leftarrow \text{lm}(Y \sim X + Z) \]

yielding a multiple regression with two predictors. The corresponding mathematical notation would be

\[ Y_i = \beta_0 + X_i\beta_1 + Z_i\beta_2 + \epsilon_i. \]

Importantly, the use of the “+” symbol in this context is different than its usual meaning; the R formula notation is just a short-hand for which variable to include in the statistical model.
and how. The following table lists the meaning of these symbols when used in an R modeling formula.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Example</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+X</td>
<td>include this variable</td>
</tr>
<tr>
<td>−</td>
<td>−X</td>
<td>delete this variable</td>
</tr>
<tr>
<td>:</td>
<td>X:Z</td>
<td>include the interaction between these variables</td>
</tr>
<tr>
<td>*</td>
<td>X*Y</td>
<td>include these variables and the interactions between them</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>^</td>
<td>(X + Z + W)^3</td>
<td>include these variables and all interactions up to three way</td>
</tr>
<tr>
<td>I</td>
<td>I(X*Z)</td>
<td>as is: include a new variable consisting of these variables multiplied</td>
</tr>
<tr>
<td>1</td>
<td>X - 1</td>
<td>intercept: delete the intercept (regress through the origin)</td>
</tr>
</tbody>
</table>

There is usually more than one way to specify the same model; the notation is not unique. For example the following three formulae are all equivalent:

\[
Y \sim X * Z * W \\
Y \sim (X + Z + W)^3
\]

each corresponding to the model

\[
Y_i = \beta_0 + X_i\beta_1 + Z_i\beta_2 + W_i\beta_3 + X_iZ_i\beta_4 + X_iW_i\beta_5 + Z_iW_i\beta_6 + X_iZ_iW_i\beta_7 + \epsilon_i.
\]

Likewise, each of these models

\[
Y \sim X + Z + W + X:Z + X:W + Z:W \\
Y \sim X * Z * W - X:Z:W \\
Y \sim (X + Z + W)^2
\]

corresponds to

\[
Y_i = \beta_0 + X_i\beta_1 + Z_i\beta_2 + W_i\beta_3 + X_iZ_i\beta_4 + X_iW_i\beta_5 + Z_iW_i\beta_6 + \epsilon_i,
\]

which differs from the previous model in that the three-way interaction has been omitted.

Finally, when using a data frame an additional time-saver is to use “.” to indicate “include all variables”. This is especially convenient when used in conjunction with the other symbols. Consider a data frame \(D\) which has columns \(Y, X, Z, \) and \(W\). Then the function call

\[
\text{fit <- lm}(Y \sim ., \text{data} = D)
\]

is equivalent to

\[
\text{fit <- lm}(Y \sim X + Z + W, \text{data} = D)
\]
Similarly,

\[
\text{fit} \leftarrow \text{lm}(Y \sim .-W, \text{data} = \text{D})
\]

is equivalent to

\[
\text{fit} \leftarrow \text{lm}(Y \sim X + Z)
\]

and

\[
\text{fit} \leftarrow \text{lm}(Y \sim .*W, \text{data} = \text{D})
\]

is equivalent to

\[
\text{fit} \leftarrow \text{lm}(Y \sim X + Z + W + X:W + Z:W)
\]

Using this notation permits a data analyst to run a spate of regression specifications without having to reconfigure the columns of a spreadsheet each time.