Functions that Demonstrate the Sampling Properties of Variable Selection

John Maindonald

October 7, 2013

The function `varselect()` in the *leaps* package can be used for variable selection. Available appraoches are *forward*, *backward* and *exhaustive* selection. The *DAAG* package has the functions `bestsetNoise()` and `bsnVaryNvar()` that are designed to give insight on the sampling properties of output from the function `lm()`, when one of these variable selection approaches has been used to choose the explanatory variables that appear in the model.

1 Selection of a specified number of explanatory variables

The function `bestsetNoise()` (*DAAG*) can be used to experiment with the behaviour of various variable selection techniques with data that is purely noise. Maindonald & Braun (2010, Section 6.5, pp. 197-198) gives examples from the use of this function. For example, try:

```
bestsetNoise(m = 100, n = 40, nvmax = 3)
bestsetNoise(m = 100, n = 40, method = "backward",
    nvmax = 3)
```

The analyses will typically yield a model that, if assessed using output from R's function `lm()`, appears to have highly (but spuriously) statistically significant explanatory power, with one or more coefficients that appear (again spuriously) significant at a level of around p=0.01 or less.

The function `bestsetNoise()` has provision to specify the model matrix. Model matrices with uncorrelated columns of independent Normal data, which is the default, are not a good match to most practical situations.

2 Change with the Number of Variables Available for Selection

As above, datasets of random normal data were created, always with 100 observations and with the number of variables varying between 3 and 50. For three variables, there was no selection, while in other cases the "best" three variables were selected, by exhaustive search. Figure 1 plots the p-values for the 3 variables that were selected against the total number of variables. The fitted line estimates the median p-value.

```
## Code
library(quantreg, quietly = TRUE)
library(splines, quietly = TRUE)
set.seed(37)  # Use to reproduce graph that is shown
bsnVaryNvar(m = 100, nvar = 3:50, nvmax = 3)
```

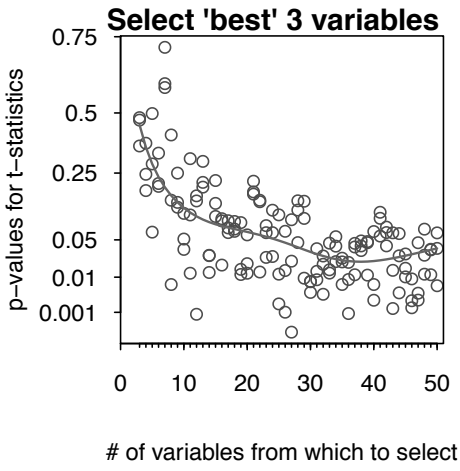

Figure 1: *p*-values from R's `lm()` function, versus number of variables available for selection, when the "best" 3 variables were selected by exhaustive search. A curve has been added that estimates the median *p*-value, as a function of `nvar`. The function `bsnVaryNvar()` that is used for the calculations makes repeated calls to `bestsetNoise()`. Similar results will be obtained from use of forward or backward selection.

When all 3 variables are taken, the *p*-values are expected to average 0.5. Notice that, for selection of the best 3 variables out of 10, the median *p*-value has reduced to about 0.1.

References

MAINDONALD, J. H. AND BRAUN, W.J. 2010. *Data Analysis and Graphics Using R – An Example-Based Approach*, 3rd edition, Cambridge University Press.
`http://www.maths.anu.edu.au/~johnm/r-book.html`