Valid p-values for stepwise regression and other post-hoc model selection methods

Richard A. Harshman and Margaret E. Lundy

University of Western Ontario, London, Ontario, Canada
harshman@uwo.ca

Abstract

When model modifications are selected using post-hoc information (e.g., in stepwise regression) standard estimates of p-values become biased. We show that this bias has two components: the advantage of getting the best current option, and the disadvantage of not getting the better alternatives chosen at prior steps. The relative impact of these two effects shifts across steps, and it is further modulated by the specific intercorrelations among the variables in and out of the model at each step. This makes the net bias extremely difficult to derive analytically. We therefore developed an alternative, compute-intensive approach that empirically determines at each step the appropriately adjusted null distribution, from which one obtains valid p-values for observed effects. To test the method, we have applied it to the improvements in fit obtained during forward selection of predictors in a stepwise regression procedure.

The method derives null-distribution values by random but synchronous row permutation of the vectors not yet entered into the model. An important new feature is the use of “null-set pruning” (elimination of cases inconsistent with prior step results). The combined result is to empirically generate the appropriate null distribution at each step. Our Monte Carlo tests to date indicate that uniform and unbiased p-values are obtained at every step. Potential applications also include valid p-values for post hoc group comparisons and stepwise canonical correlation (and hence ANOVA/MANOVA/Discriminant Analysis, etc.).

1. The Motivation: a need to choose

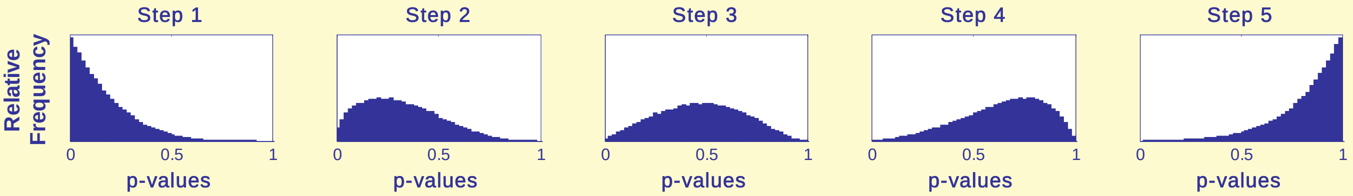
When faced with too many alternatives (e.g., too many potential predictors for a multiple regression model), one needs to make tough choices. This is sometimes done by starting with the simplest model and then incrementally adding model improvements, each time choosing the 'best' addition from among a set of remaining possibilities.

Which is **'best' is determined by post hoc comparison** of their relative statistical effects. (Examples of this include adding predictors in multiple regression by forward selection, or adding group contrasts during post hoc multiple comparisons of group means).

Naturally, we **want to know the statistical significance of each improvement** -- the likelihood of obtaining such an improvement by chance under a null hypothesis that the increment is due to chance.

2. The problem: choice introduces bias

Example 1 ('parametric' p-value estimates in a 5 steps procedure):



These histograms show the relative frequency of p-values of different sizes that were found by a standard F-test when applied in a forward selection Multiple Regression procedure.

100,000 simulated cases were analyzed. Each consisted of a random **y** and 5 random **x** vectors (vector elements were drawn from N(0,1) then centered). At each step the vector making the best improvement was entered into the model, until all 5 vectors were entered (Step 5).

At Step 1, unrealistically low p-values were reported. At successive steps, the distribution shifted toward larger values. An interplay of positive and negative bias in the middle step suppressed both large and small p's. Since all values were random, **an unbiased estimate of p-values should have a uniform distribution.**

Of course, we did not discover the problem of bias caused by model selection. Sources, .e.g., in our references discuss the problem clearly.

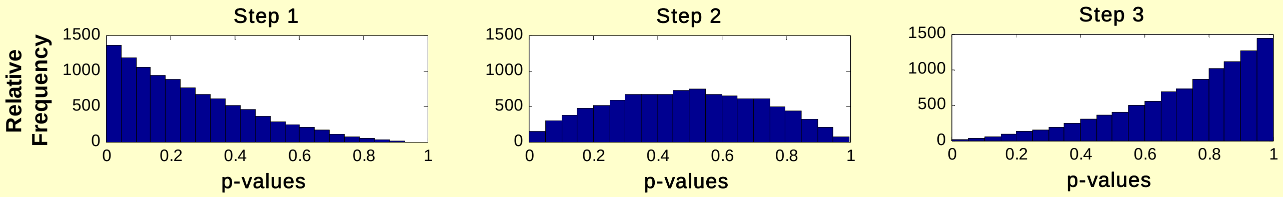
3. Analyzing the Bias: two shifting aspects

The same shifting bias is seen when p-values are estimated ‘nonparametrically’ by a permutation test. In this case the data consist of a random y and only 3 random x vectors, so there are only three potential predictors in the selection set.

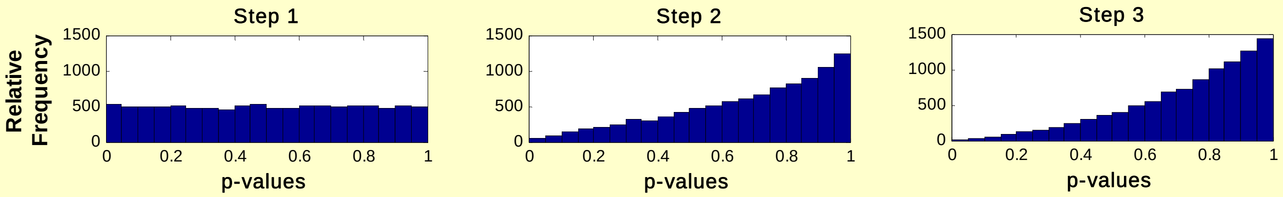
The bias shifts because of a changing relative contribution of two components: (i) the advantage of getting the best current option, and (ii) the disadvantage of not getting the even better alternatives chosen at prior steps.

Example 2 (‘nonparametric’ p-values, 3 regression steps):

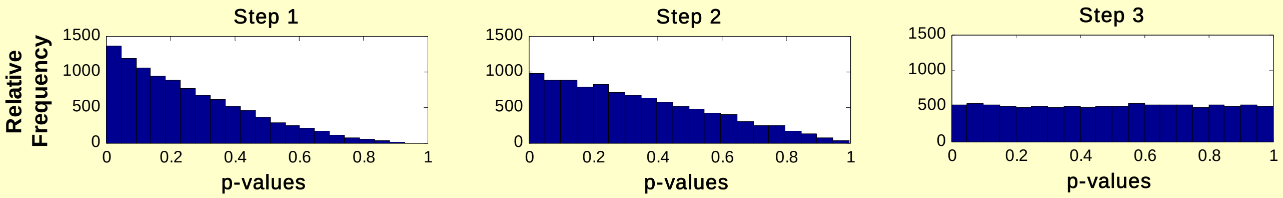
Case 2A: Without any bias correction, we see a shifting bias similar to Case 1:



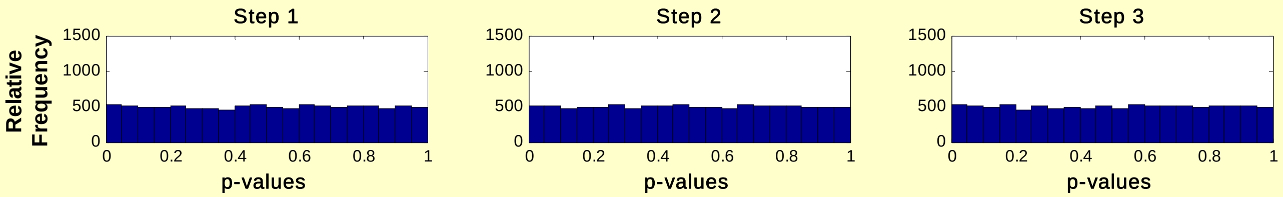
Case 2B: Below, we remove the ***current-choice advantage*** by incorporating the same choice advantage into the generation of null set values. No bias is left on Step 1, but the later steps still show an increasing *prior choice disadvantage:*



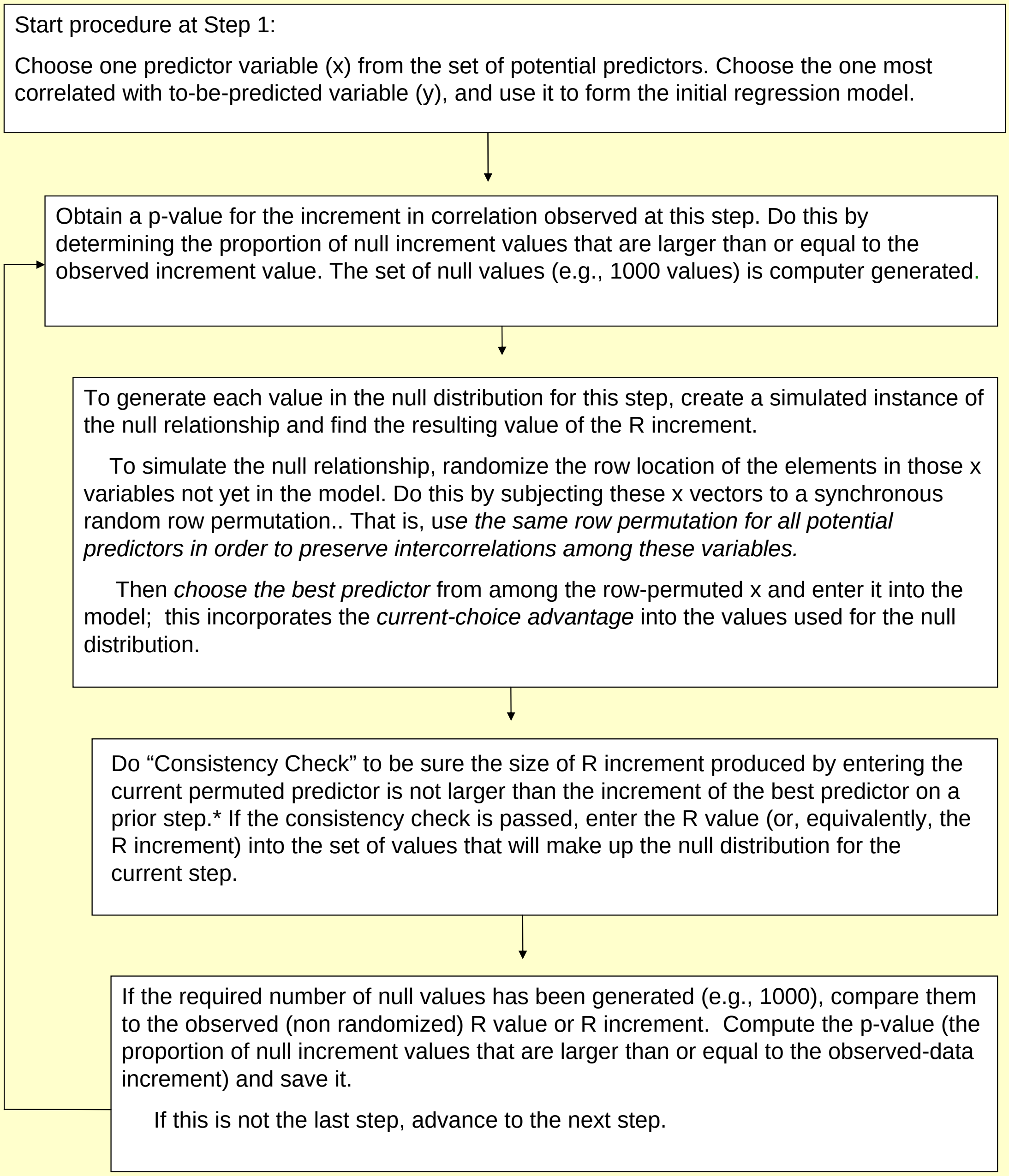
Case 2C: In contrast, we below remove instead the ***prior-choice disadvantage*** by applying “null set pruning” to eliminate from the null distribution spurious large prediction increments that are larger than the best ones found at earlier stages:



Case 2D: By using both corrections when generating null set values, unbiased p-values are obtained at all steps:



4. Algorithm:



Start procedure at Step 1:

Choose one predictor variable (x) from the set of potential predictors. Choose the one most correlated with to-be-predicted variable (y), and use it to form the initial regression model.

Obtain a p-value for the increment in correlation observed at this step. Do this by determining the proportion of null increment values that are larger than or equal to the observed increment value. The set of null values (e.g., 1000 values) is computer generated.

To generate each value in the null distribution for this step, create a simulated instance of the null relationship and find the resulting value of the R increment.

To simulate the null relationship, randomize the row location of the elements in those x variables not yet in the model. Do this by subjecting these x vectors to a synchronous random row permutation.. That is, u*se the same row permutation for all potential predictors in order to preserve intercorrelations among these variables.*

Then *choose the best predictor* from among the row-permuted x and enter it into the model; this incorporates the *current-choice advantage* into the values used for the null distribution.

Do “Consistency Check” to be sure the size of R increment produced by entering the current permuted predictor is not larger than the increment of the best predictor on a prior step.* If the consistency check is passed, enter the R value (or, equivalently, the R increment) into the set of values that will make up the null distribution for the current step.

If the required number of null values has been generated (e.g., 1000), compare them to the observed (non randomized) R value or R increment. Compute the p-value (the proportion of null increment values that are larger than or equal to the observed-data increment) and save it.

If this is not the last step, advance to the next step.

* if 5000 consecutive null r’s have to be eliminated, the dataset is abandoned

5. Summary of the Method: four nested loops

The procedure can be roughly summarized as a series of nested loops:

REPLICATION LOOP – generate random data (1 million sets, X=20x10, y=20x1)

REGRESSION LOOP – do 3 steps

STEP LOOP – get p-value for the step

PERMUTATION LOOP – get null r distribution (1000 values)
from which the p-value is computed

CONSISTENCY CHECK LOOP – eliminate "inconsistent" null r's
this is done for all but the first step

If 5000 consecutive null r's have to be eliminated, the dataset is abandoned.

6. Testing the method: Monte Carlo simulation

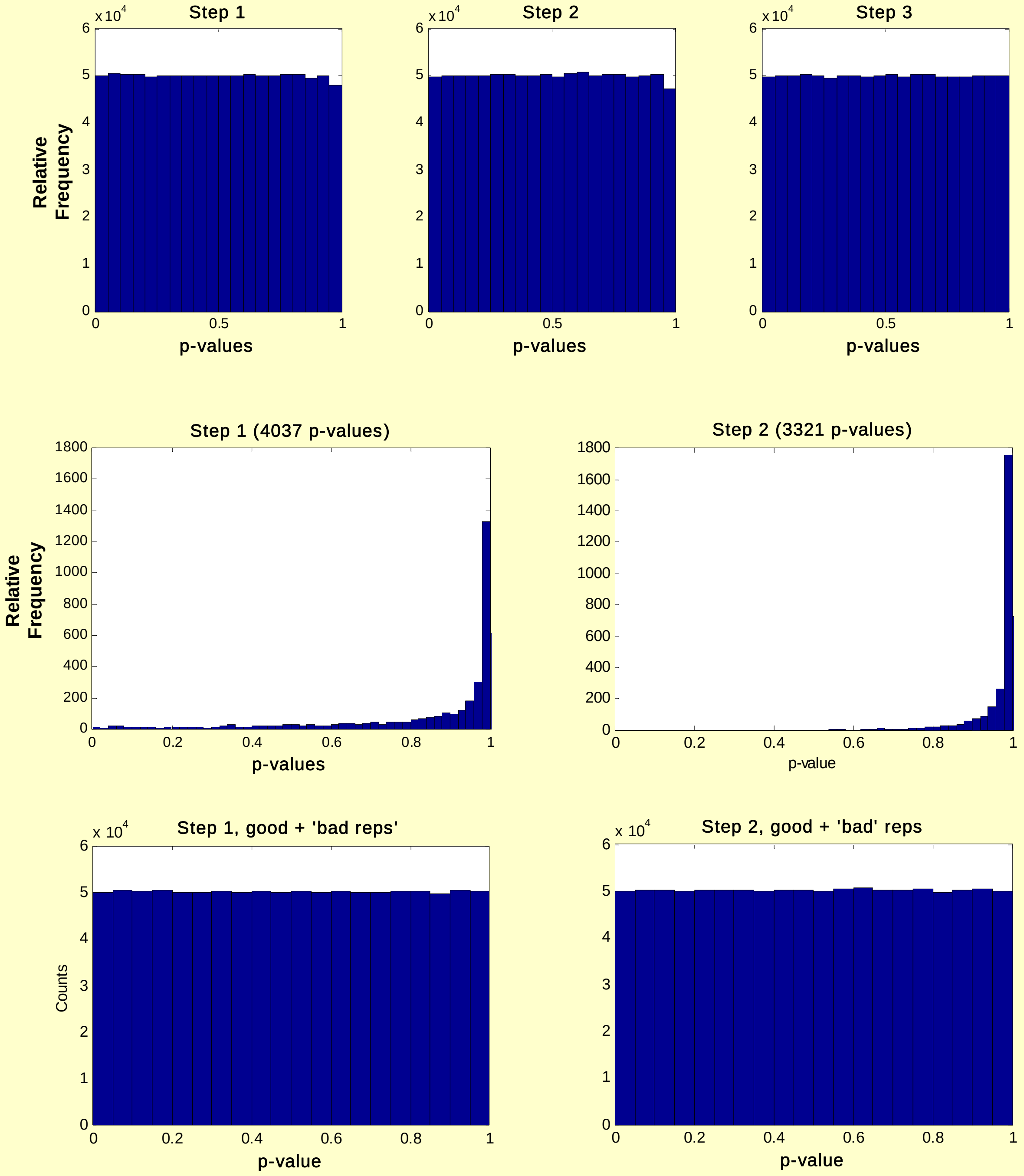
A series of Monte Carlo tests were performed, evaluating the distribution of p-values obtained by different methods when applied to random data (i.e., in the null situation). The resulting distributions were examined/tested for closeness to a uniform distribution, which, if found, would indicate absence of bias.

No "stopping criterion" was used. Instead, a specified number of steps was performed regardless of the p-value at any particular step.

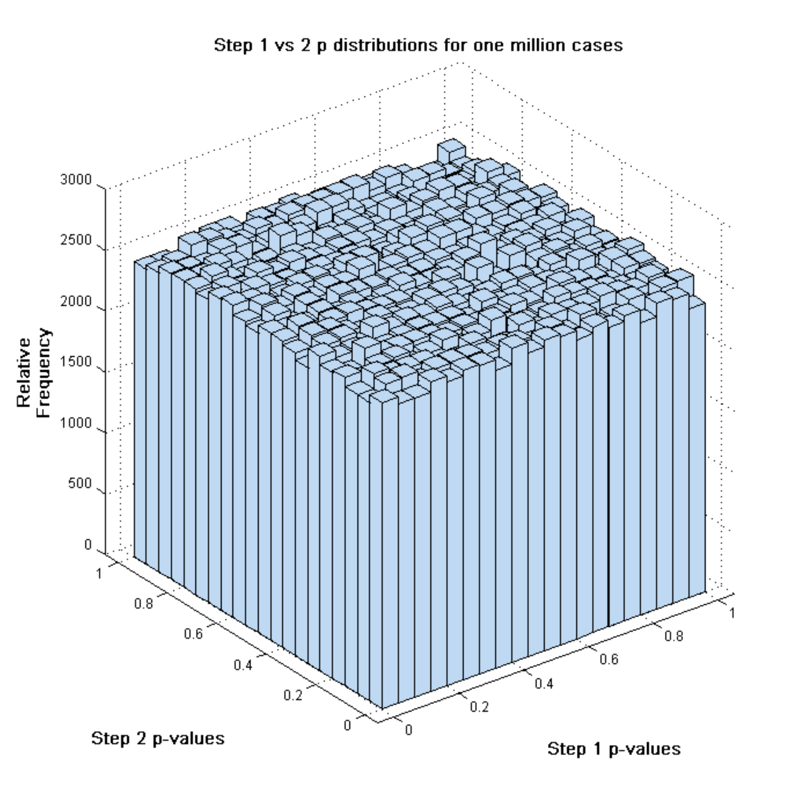
Naturally, some of these random cases happened to include one or more high predictive relations; these cases were examined to be sure the method provides unbiased p-values even following a strong predictive increment. Performance was found to be consistent even in these cases, as indicated by the two-way histogram shown at the right.

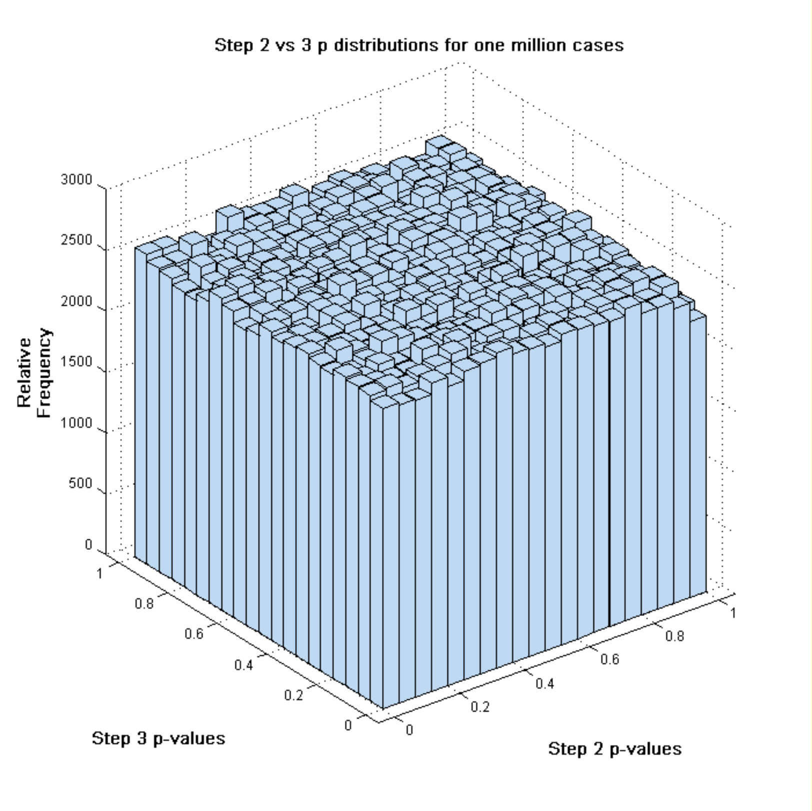
7. Results: no bias detected

The following histograms are based on one million multiple regressions using forward selection to construct 1, 2, and 3 predictor models from completely random data. The p-values estimated by our proposed method appear uniform at all steps (except the very rightmost interval, which is an artifact of the rejection of datasets when consistent null values could not be found on subsequent steps after 5000 successive attempts). In the second row histograms below, the number of these aborted cases is shown. Since these are rejected because of problems on subsequent steps, we can add the values back into Step 1 and 2, resulting in the histograms in the third row below. The “notch” has vanished and the procedure looks uniform for all p-value.



Joint distribution shows independence across steps





8. Conclusion/Discussion: many applications

A working method with many applications

So far, our Monte Carlo tests indicate that this method is either unbiased or so slightly biased that the bias cannot be detected with one million cases. This suggests that the method, and variants of it, might be useful in a wide range of analyses where post hoc information is used – or could usefully be introduced.

Because the method is built upon a theoretical account of how the bias arises and can be corrected, variations of the method consistent with this theory seem likely to be successful.

For example, the method can be generalized to introduce model/data selection into such procedures as MANOVA and Discriminant Analysis, by implementing an incremental selection logic in canonical correlation (in fact, such a procedure is in development and is already partly programmed).

Post Hoc tests

In collaboration with Dr. R. C. Gardner, also at the University of Western Ontario, we have begun to explore application of the method to post hoc tests and the problem of multiple comparisons for ANOVA , chi-square, etc. One approach incrementally predicts dependent variables by contrast-coded vectors..

Acknowledgements

Research funded by the Natural Sciences and Engineering Research Council of Canada (RGPINs 7896-01 and -06)

Some of this work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca)

References

Edgington, E. S. (1986). Randomization tests (2nd ed.). New York: Marcel Dekker, Inc.

Good, P. (2000). Permutation tests (2nd ed.). New York: Springer.

Hjorth, J. S. U. (1994). Computer intensive statistical methods: Validation model selection and bootstrap. New York: Chapman & Hall.

Grechanovsky, E., & Pinsker, I. (1995). Conditional p-values for the F-statistic in a forward selection procedure. Computational Statistics & Data Analysis, 20, 239-263.