A randomization method of obtaining valid p-values for model changes selected "post hoc"

Richard A. Harshman and Margaret E. Lundy
University of Western Ontario, London, Ontario, Canada
Psychometric Society Annual Meeting June, 2006, HEC, Montreal, CANADA,
harshman@uwo.ca, mlundy@uwo.ca

Abstract

When model changes are guided by *post hoc* assessment of the observed improvements in prediction or fit (e. g. , in stepwise regression), it is usually impossible to obtain p-values for these improvements by conventional analytical methods. We describe a computer-intensive alternative that accurately estimates these p-values by using a modified randomization / permutation test procedure that empirically determines the appropriate null distributions. To demonstrate the method, we use it to get valid p-values for step improvements found during standard stepwise multiple regression. Our method corrects for bias caused by the increased 'capitalization on chance' intrinsic to post hoc variable selection; it does this by introducing an equivalent post hoc selection step into the process generating the null-hypothesis values. The method also corrects for an "inconsistency" bias by eliminating or "pruning out" permuted cases that are inconsistent with prior step results; without such pruning, the method would underestimate significance except on the first step. In a Monte Carlo sample of one million cases, the p-values estimated for fit improvements during a three-step stepwise multiple regression did not show a statistically detectable bias at any step. Potential applications include significance tests for more complex sequential methods, stepwise canonical correlation / MANOVA, and discriminant analysis.

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 sometimes starts with the simplest model and then incrementally adds model improvements, each time choosing the 'best' addition from among a set of remaining possibilities.

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

Example: Rating treatments for back pain

In 2002 we collaborated with Drs. L. Swartzman, J. Burkell, and others at UWO in an investigation of peoples' perceptions of "alternative" vs. "conventional" medical treatments (hypothetically, for chronic back pain).

In the data reported here, 89 participants rated each back treatment on 20 objective properties (e.g., 'Painful'; 'Very well researched'; 'Has serious side effects') and also on 9 evaluative scales (e.g., 'A good approach'; 'Makes me uneasy to even <u>think</u> about it'; 'A really dumb thing to try', etc.).

One approach to studying the objective-subjective connections is to use multiple regression to predict *evaluative* ratings for a given treatment from the *objective* ratings of that treatment.

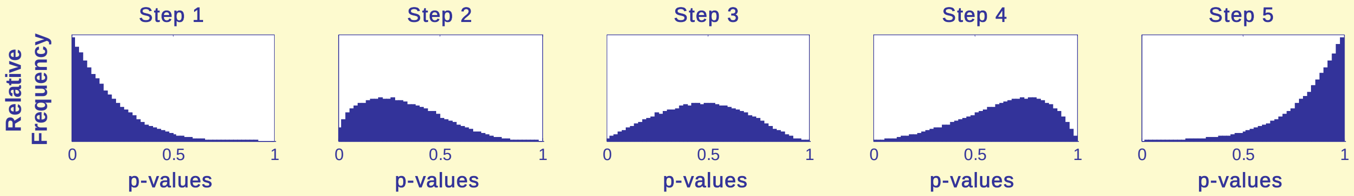
Here are the 20 objective properties rated on 9 point scales
(the potential predictors):

1. Requires effort from the person treated
2. Painful
3. Very invasive
4. Very effective
5. Very well researched
6. Requires effort from the health care provider
7. Natural approach
8. Has serious side effects
9. Treats the body
10. Disruptive to one's daily routine
11. Very dangerous
12. Under the control of the person treated
13. Treats the mind
14. Conventional medical treatment
15. Very harsh
16. Holistic
17. Foreign to the body
18. Affects the whole body
19. We completely understand how this treatment/management approach works
20. Very "penetrating"

With 20 potential predictors and only 89 raters, we could not simultaneously enter all into a regression model. Instead, we entered them one at a time, using a forward selection procedure, and evaluated the p-value of the improvement in R at each step (until p>.5).

2. The problem: choice introduces bias

Example 1 ('parametric' p-value estimates in a 5 step 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.**

Prior work

We are, of course, not the first to point out that bias is introduced by model selection. The invalidity of standard p-value computations for stepwise regression is well known. In the broader context of model selection generally, recent authors such as Hjorth (1994) have discussed the problem extensively, and provided alternative strategies, but not a way to obtain correct p-values.

However, in 1995, Grechanovsky & Pinsker published an algebraic derivation of valid p-values for the F-test in forward selection multiple regression. Their treatment is mathematically sophisticated and uses the standard distributional assumptions. The method described here grows out of a more general theory of bias, described below, and an empirical approach that is approximate, computationally intensive, but at the same time nonparametric and very flexible – potentially adaptable to many different kinds/applications of model selection.

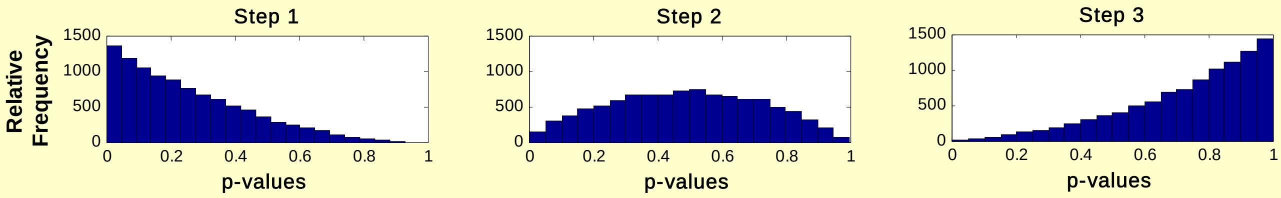
3. Analyzing the Bias: two shifting aspects

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.

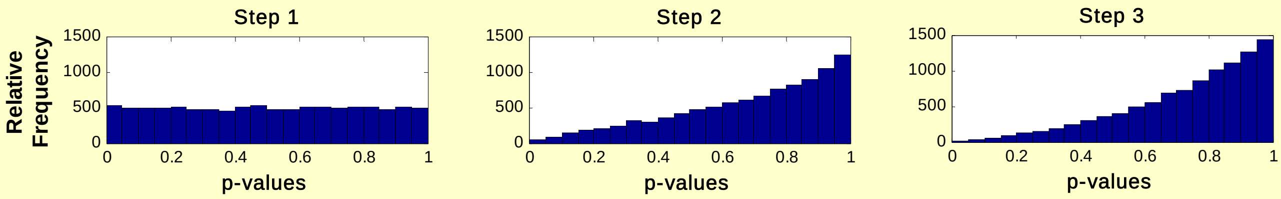
The bias shifts because of a changing relative contribution of two components: (i) the advantage of getting the best current option, and (ii) the disadvantage of not getting the even better alternatives chosen at prior steps.

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

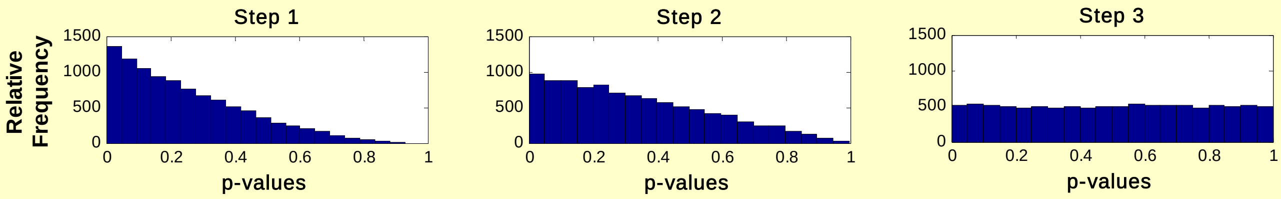
Case 2A: Without any bias correction, we see a shifting bias similar to Case 1:



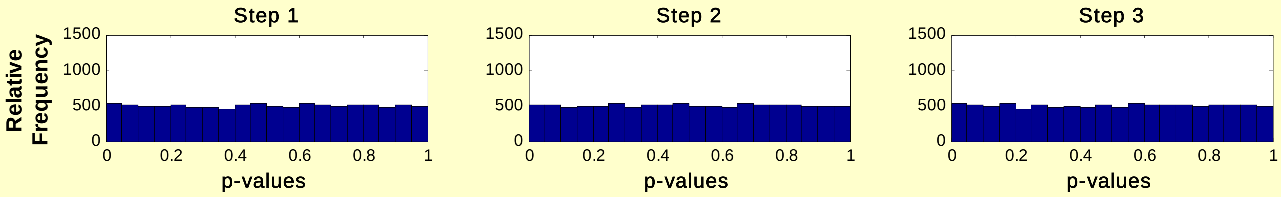
Case 2B: Below, we remove the ***current-choice advantage*** by incorporating the same choice advantage into the generation of null set values. No bias is left on Step 1, but the later steps still show an increasing *prior choice disadvantage:*



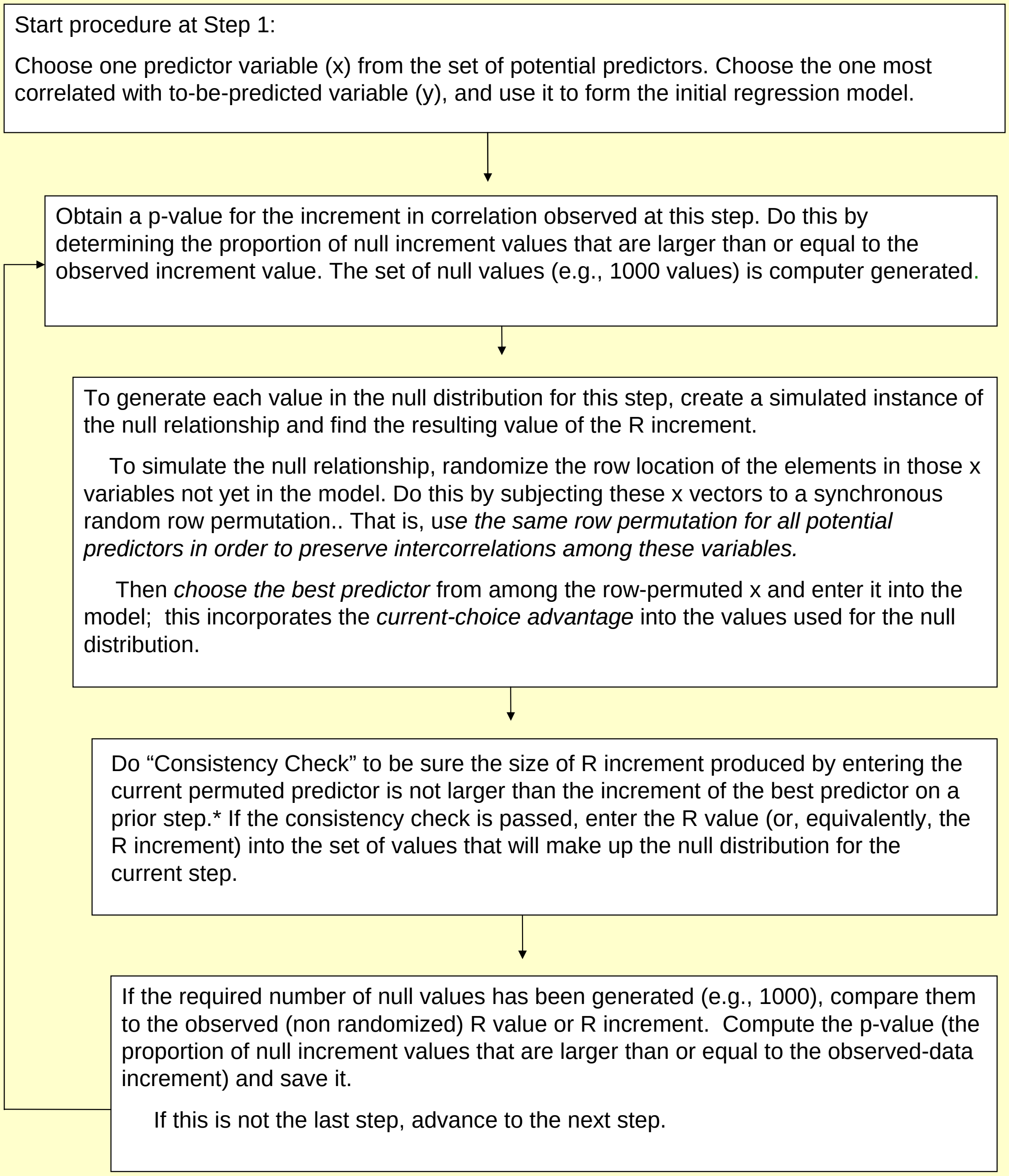
Case 2C: In contrast, we below remove instead the ***prior-choice disadvantage*** by applying “null set pruning” to eliminate from the null distribution spurious large prediction increments that are larger than the best ones found at earlier stages:



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



4. Algorithm: picking and pruning



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

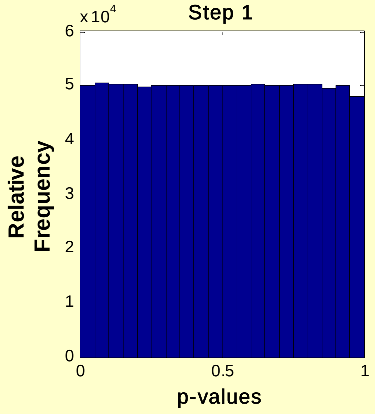
A series of Monte Carlo tests were performed, evaluating the distribution of p-values obtained by different methods when applied to random data (i.e., in the null situation). The resulting distributions were examined/tested for closeness to a uniform distribution, which, if found, would indicate absence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

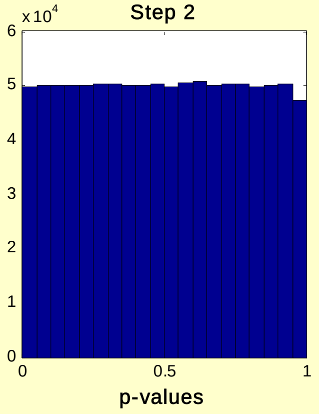
No "stopping criterion" was used. Instead, a specified number of steps was performed regardless of the p-value at any particular step.

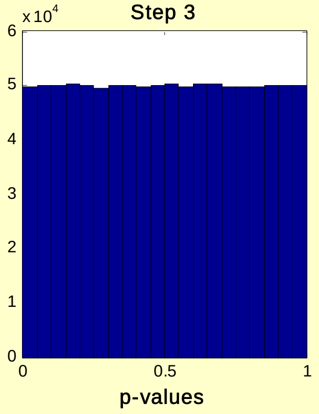
Naturally, some of these random cases happened to include one or more high predictive relations; these cases were examined to be sure the method provided unbiased p-values even following a strong predictive increment. Performance was found to be consistent even in these cases, as indicated by the two-way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

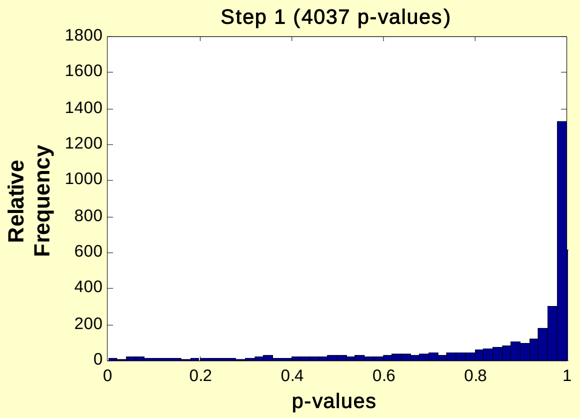
7. Results: no bias detected

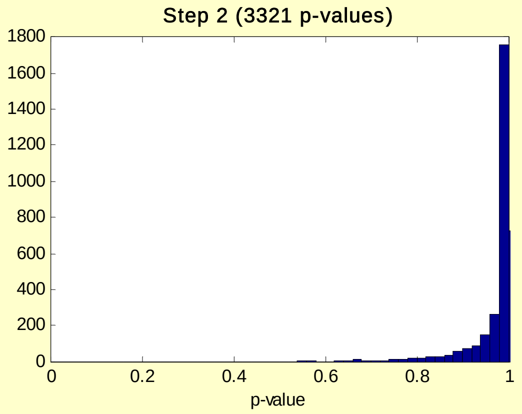
The following histograms are based on one million multiple regressions using forward selection to construct 1, 2, and 3 predictor models from completely random data. The p-values estimated by our proposed method appear uniform at all steps (except the very rightmost interval, which is an artifact of 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-values.

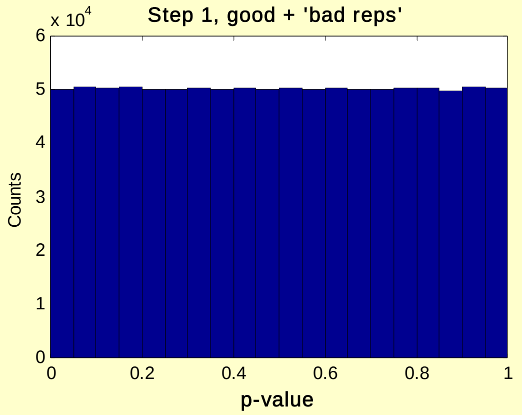


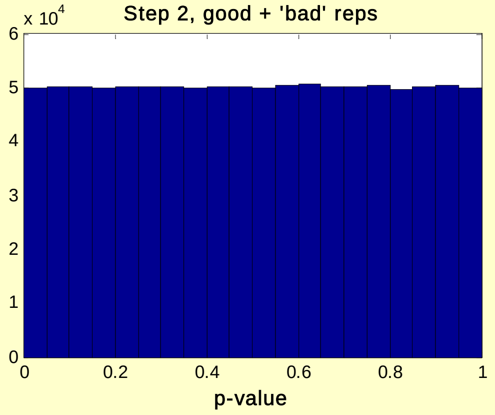




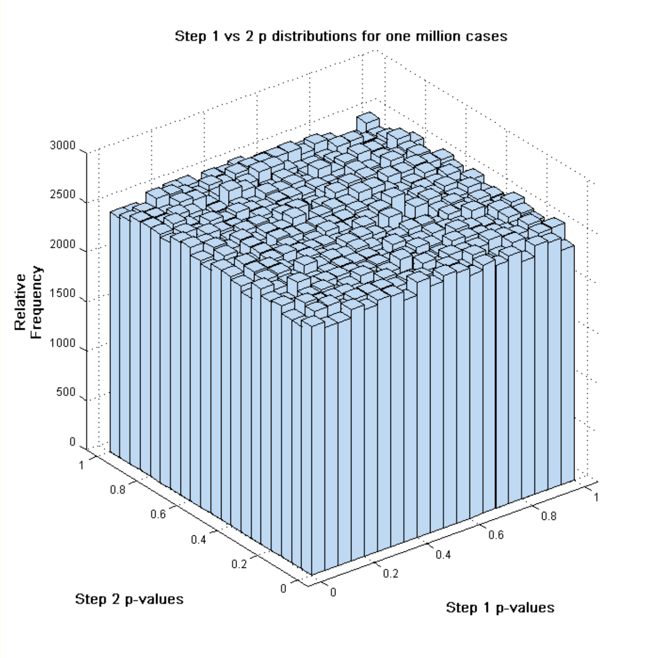


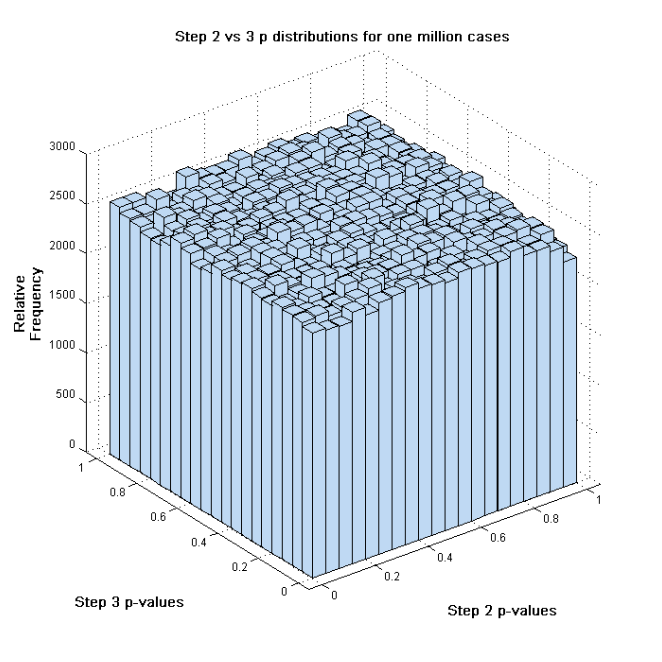






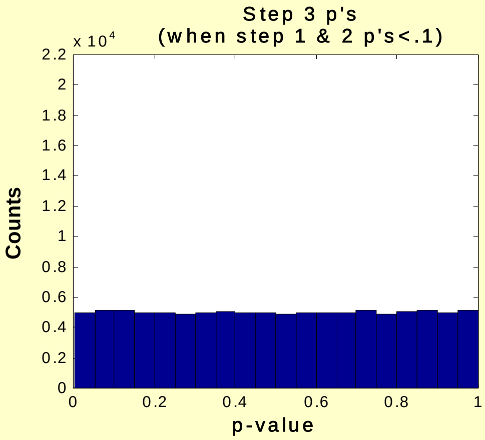
Joint distribution shows independence across steps

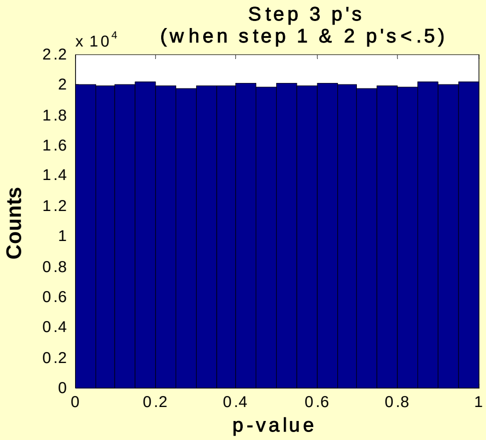




Conditional stepping also unbiased

In the Monte Carlo results reported above, the procedure performed all three steps regardless of the p at each step. In most applications, however, one goes to the next step only when the current one meets some criterion (typically, p-value < threshold). To simulate this, we extracted all Step 3 p's where the prior two steps had p<.1 (left) or p<.5 (right). These also showed no indication of bias.





Example: Rating treatments for back pain

Below are some regression results demonstrating the differences in p-values that result from elimination of selection bias (in this case, bias due to selection from 20 possible predictors):

Treatment: **Back surgery (requiring hospital stay)**
Liking scale to be predicted: **A good approach [for treating chronic back pain]**

Regression	p-values		B-weight	Multiple	
Step	Ours	Uncorrected	sign	r	Predictor (objective scale)

1	.1706	.0101	+	.270	Under patient control
2	.1810	.0213	-	.358	Has serious side effects

3	.4859	.0609	+	.405	Very well researched
4		.2125	-		Disruptive to daily routine
5		.1652	+		Very invasive
6		.1784	-		Very “penetrating”

Treatment: **Chiropractic Adjustments**
Liking scale to be predicted: **A good approach [for treating chronic back pain]**

Regression	p-values		B-weight	Multiple	
Step	Ours	Uncorrected	sign	r	Predictor (objective scale)
1	.0001	.0000	+	.588	Very effective

2	.1562	.0095	-	.629	Very dangerous
3	.2434	.0261	-	.656	Very “penetrating”

4		.0857	+		We understand how it works
5		.0659	+		Foreign to the body
6		.1277	-		Disruptive to daily routine

Treatment: **Injections of pain killing medication into affected area**
Liking scale to be predicted: **Makes me uneasy to even think about it [for treating chronic back pain]**

Regression	p-values		B-weight	Multiple	
Step	Ours	Uncorrected	sign	r	Predictor (objective scale)
1	.0004	.0000	+	.452	Painful

2	.2040	.0131	+	.510	Has serious side effects
3	.2250	.0279	+	.549	Very well researched
4	.3766	.0575	-	.575	Very effective
5	.0041	.0205	-	.610	Treats the mind

6	.4861	.0970	+	.627	Requires patient effort

Treatment: **Non-narcotic prescription pain killing tablets**

Liking scale to be predicted: **A good approach [for treating chronic back pain]**

Regression	p-values		B-weight	Multiple	
Step	Ours	Uncorrected	sign	r	Predictor (objective scale)
1	.0025	.0001	+	.394	Natural approach
2	.0189	.0011	+	.505	Conventional treatment

3	.0740	.0052	+	.566	Very effective
4	.2525	.0229	-	.601	Has serious side effects

5		.1318	+		Treats the body
6		.1367	+		Very harsh
7		.1668	-		Very well researched
8		.2254	+		Under patient control

Treatment: **Herbal teas**

Liking scale to be predicted: **A really cool thing to try [for treating chronic back pain]**

Regression	p-values		B-weight	Multiple	
Step	Ours	Uncorrected	sign	r	Predictor (objective scale)
1	.0308	.0021	+	.322	We understand how it works
2	.1068	.0085	+	.417	Treats the body
3	.0009	.0016	-	.515	Foreign to the body

4	.1077	.0126	+	.564	Is disruptive to daily routine

5		.1507	+		Conventional medical treatment
6		.1536	-		Very invasive
7		.1112	+		Natural approach
8		.1507	+		Very “penetrating”

Approximation accuracy

Because each null distribution is approximated by a random sequence, there will be some variation from one run to the next. Below are three different sets of p-values for the same multiple Rs (from ratings of ‘Herbal Teas’).

In this case, 10,000 random null cases were generated each time, requiring more than a minute on a typical Pentium desktop computer. The ‘jitter’ in the estimates is roughly .005.

Step	Mult. R	Seed 1	Seed 2	Seed 3
1	.322	.0308	.0384	.0361
2	.417	.1068	.1054	.1112
3	.515	.0009	.0011	.0006
4	.564	.1077	.1116	.1101

8. Discussion/Conclusion:

a working method with many applications

This empirical method, which adjusts the null for selection bias while pruning out invalid null cases might be called 'pick and prune'. Our application indicates that compact and sensible relationships can be recognized once the serious bias and inflated significance due to selection is removed. 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.