Stepwise vs. Block-Exhaustive Regression
A Comparison using Clinical Data

Antony Davies, PhD
June 1, 2001

PARABON™ COMPUTATION | Computing Outside the Box℠

Introduction

This report compares the results of two analyses of clinical data using 70 potential explanatory factors across 110 patients in an attempt to find the model that best describes the patient outcome variable.[1] Without recourse to an underlying theoretical model, this data was analyzed using traditional stepwise regression and using *Exhaustive Regression*, statistical analysis application software designed to leverage the massive computational resources of the Frontier™ distributed computing platform.

Employing massive computational power, Exhaustive Regression is designed to look at all possible linear models of up to 35 to 40 explanatory factors and report those for which all parameter estimates are statistically significant. In this case, 70 regressors implied a search space of 10^{21} linear combinations. To exhaustively search this space requires orders of magnitude more computing power than exists worldwide.

To make the problem manageable, Exhaustive Regression was applied to a sequence of subsets of the data to achieve "block-exhaustive" regression. In block-exhaustive regression, a subset of the full data set is exhaustively searched. Factors that show minimal effect on the dependent variable are removed from the data set and replaced with factors that had been held out. This new subset is then exhaustively searched. Block-exhaustive regression continues in this fashion until all of the data has been examined.

For this data set, we performed a sequence of five Exhaustive Regression runs (each with 25 to 30 regressors). Within a single Exhaustive Regression run, only those factors that failed to appear in any of the statistically significant models were eliminated from further consideration.[2] In total, Exhaustive Regression examined 2.2 billion linear models. The analysis, which would have required over a year to complete on a single computer, finished in 15 hours.

Results

Of the 2.2 billion models examined, there were 249 models in which all parameter estimates were statistically significant. Of the 70 factors, only 25 appeared in at least one of the significant models. The goodness of fit (adjusted squared multiple correlation) for these models ranged from 0.09 to 0.27.

---

[1] Variable names are obfuscated to protect the client's intellectual property.
[2] Here, a model is considered statistically significant if all of the parameter estimates are significant at the $\alpha = 0.10$ level. For example, in a run of 25 regressors, Exhaustive Regression performs all 33.6 million possible linear regressions. Each regression model in which all parameter estimates are significant at the 10% level is returned. If a given factor fails to appear in any of these returned results, then that factor is removed from the data set for the next Exhaustive Regression run.

In evaluating factors across models, we employ a factor evaluation heuristic (H) that is calculated as follows:

$$H = \left| \ln\left(1 - \bar{R}^2\right) \right|$$

Each model yields a value for H. For each model, this value is assigned to all statistically significant regressors that appear in the model. This heuristic is superior to the frequency of factor appearance because the heuristic adjusts for the goodness-of-fit of the different models. It is also superior to simply summing the $\bar{R}^2$ for models in which a factor appears because it weights incremental changes in $\bar{R}^2$ more heavily as $\bar{R}^2$ approaches 0 or 1. It is important to note that, while H is a heuristic, its statistical properties are not known.

One of the problems with a statistical search algorithm is the likelihood of finding relationships that are, by mere random chance, statistically significant. Ranking models by the summed heuristic partially addresses this issue by weighting model selection in favor of regressors that consistently appear in well-fitted models, regardless of what other regressors are present. Because of the likelihood of finding spurious results, the results of any search algorithm should be treated as the launch-point for hypothesis formation and further experimentation, not as the definitive conclusion of a study.

The twenty-five regressors that appeared in at least one of the 249 significant models are detailed in Table 1.

Table 1. Factors Appearing in at Least One Significant Model

Factor	*H*	Mean Coefficient	Std Deviation[3]	Frequency
a	59.4	0.51	0.04	132
b	56.4	0.23	0.04	149
c	39.3	0.35	0.04	56
d	37.8	0.23	0.02	49
e	36.5	-13.86	1.55	31
f	29.3	-0.63	0.08	164
g	23.4	-0.24	0.02	138
h	20.3	9.96	0.94	112
i	18.3	22.31	2.17	102
j	17.7	-2.48	0.46	99
k	13.6	-9.38	0.86	83
l	13.5	-6.36	0.66	70
m	13.2	-26.31	2.75	67
n	10.8	1.87	0.13	59
o	5.9	-7.18	0.77	37

[3] These are the standard deviations of the parameter coefficients across models, not the mean of the standard deviations of estimates within models.

Table 1. (continued)

Factor	*H*	Mean Coefficient	Std Deviation[4]	Frequency
p	3.5	-0.50	0.07	21
q	3.3	-9.70	0.86	27
r	2.2	9.01	0.50	18
s	1.9	8.02	0.36	9
t	1.2	-14.77	0.93	4
u	1.1	-18.48	0.87	4
v	0.6	-17.22	0.08	2
w	0.6	-19.24	0.40	2
x	0.5	-43.47	0.08	3
y	0.1	8.82	N/A	1

The chart points (according to *H*) to factors that have the greatest likelihood of influence on the dependent variable assuming an unspecified linear relationship. It is likely that some factors behave as "substitute" factors (wherein two factors have strong effects independently, but the strengths are diluted when they appear together) while others behave as "complement" factors (wherein two factors have stronger effects when they appear together). To discern actions between factors, we return to the set of 249 significant models and rank these models according to $\bar{R}^2$. Results are shown in Table 2.

Table 2. Models Ranked by Goodness of Fit

Model	Factors	$\bar{R}^2$
A	a, b, c, d, e, f, g, i, j, u	0.269
B	a, b, e, e, j, s, t, u, v	0.261
C	a, b, c, d, e, f, g, j, k	0.259
D	a, b, e, i, j, s, t, v, w	0.258
E	a, b, e, f, g, i, j, k, w	0.254

An alternative scheme for model selection is to focus on models that contain factors with the greatest *H* values. Adding the *H* values for factors that appear in each model, we obtain the model rankings in Table 3.

Table 3. Models Ranked by Summed Heuristic

Model	Factors	$\bar{R}^2$
A	a, b, c, d, e, f, g, i, j, u	0.269
F	a, b, c, d, e, f, g, i, k	0.248
B	a, b, e, e, j, s, t, u, v	0.259
G	a, b, c, d, f, g, i, j, k	0.250
H	a, b, c, d, f, g, o	0.202

The models in Table 2 reflect sets of factors in which (1) all factors are statistically significant when grouped together, and (2) the set of factors, as a group, best explain

[4] These are the standard deviations of the parameter coefficients across models, not the mean of the standard deviations of estimates within models.

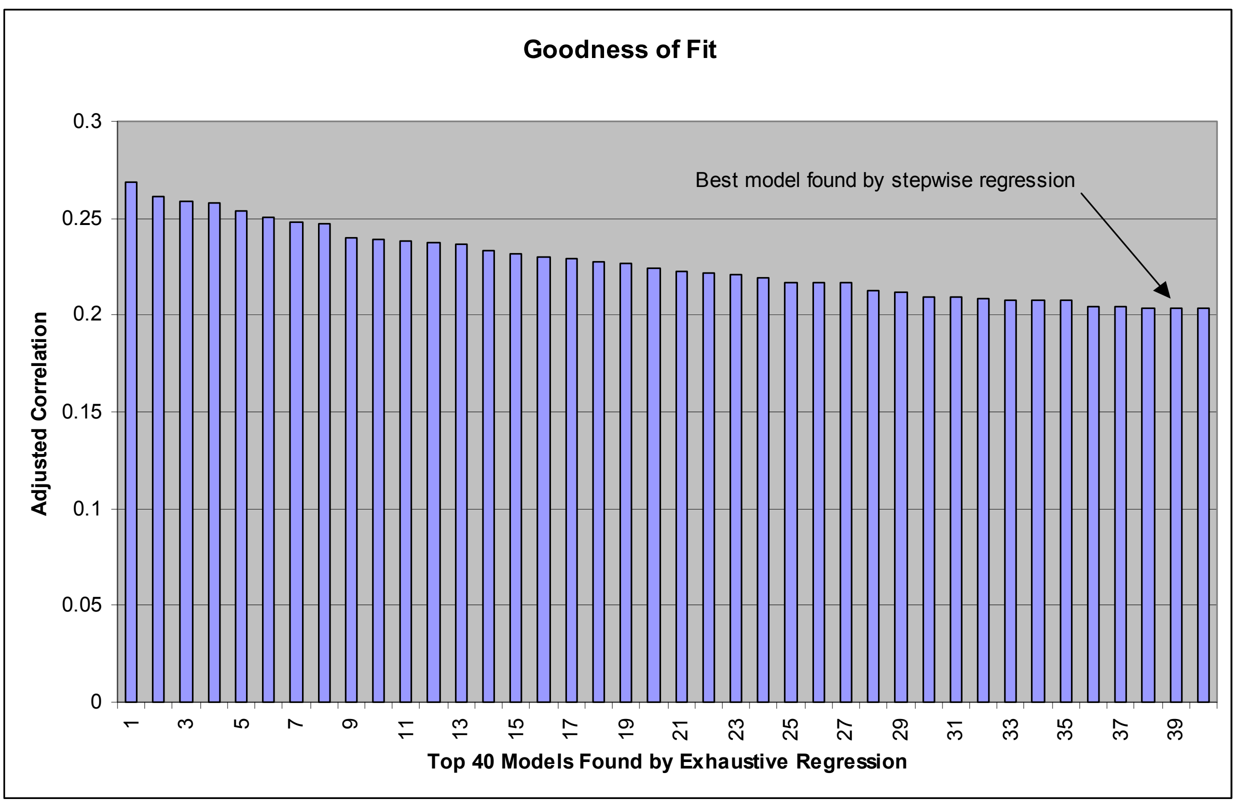
variations in the dependent variable. The models in Table 3 reflect sets of factors in which (1) all factors are statistically significant when grouped together, and (2) the total of the heuristic measures for the factors is greatest. Notice that models A and B appear among the top five models when models are ranked by either goodness-of-fit or summed heuristic, and that model A has both the greatest $\overline{R}^2$ and the greatest summed *H*.

These analyses suggest that future studies should focus on the effects of a, b, c, d, e, f, g, i, and j, on the outcome variable. Other variables with significant *H* values should also be considered in light of expected results.

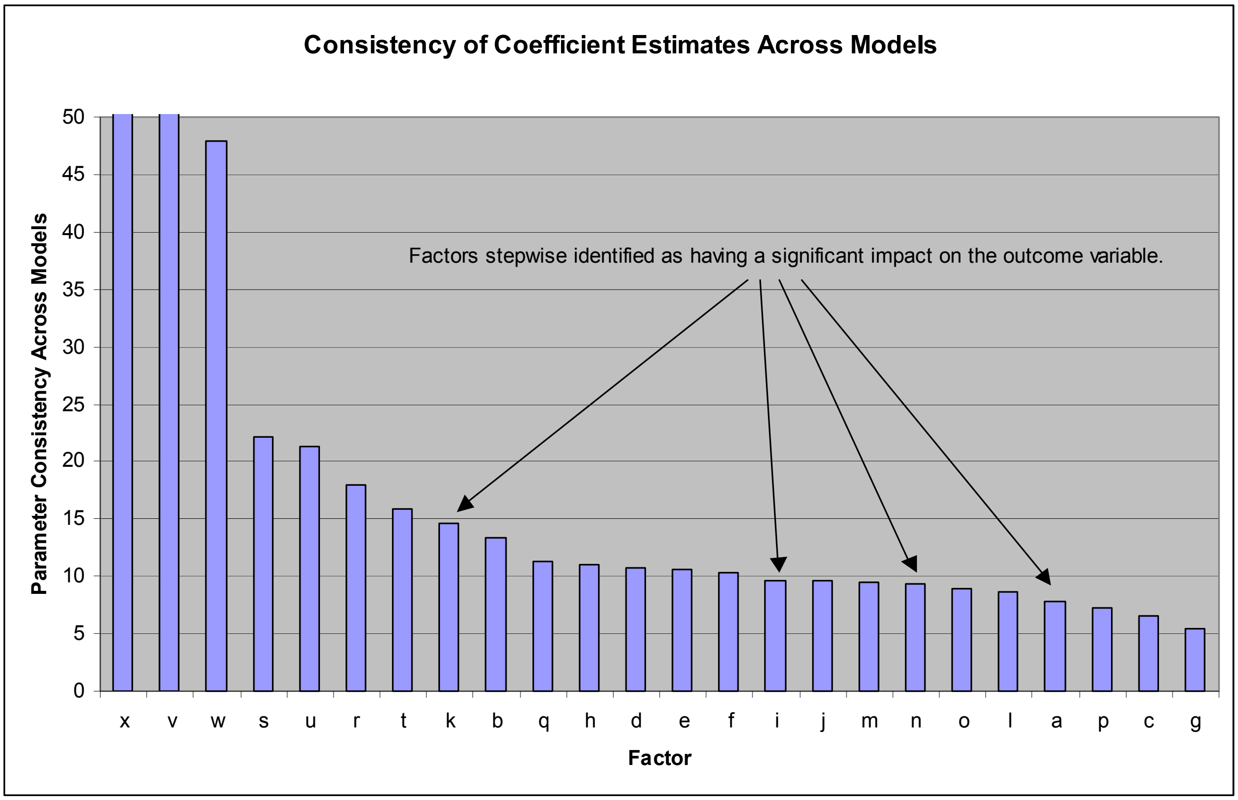
Comparison to Stepwise Regression

Stepwise regression is the traditional approach used when fitting data for which a theoretical model does not exist. Unlike Exhaustive Regression, which examines every possible linear model, stepwise regression samples only a very small set of models, swapping out variables in an attempt to find "smartly" the best combination of factors to employ in the model.

Of the 249 significant models found by Exhaustive Regression, thirty-nine were superior to the single model found by stepwise regression. The chart below shows the goodness-of-fit (as measured by adjusted squared multiple correlation) for the top forty models found by Exhaustive Regression. Note that the model found by stepwise regression ranks 39th in terms of goodness-of-fit.

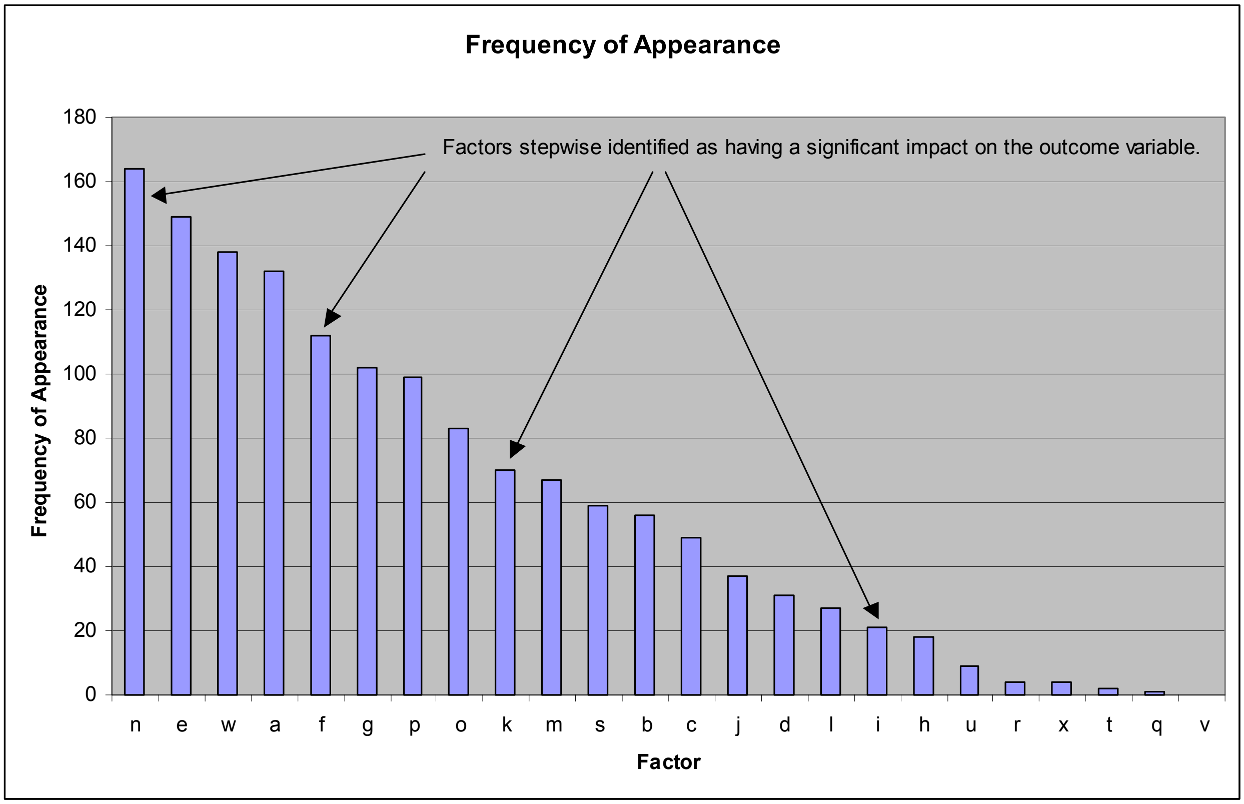


Beyond the ability simply to find better fitting models, Exhaustive Regression allows comparison of the parameter estimates across models – such comparison is impossible with stepwise because stepwise searches only a very small portion of the space of possible solutions. Factors that truly impact the dependent variable should exhibit relatively consistent parameter estimates across models. The chart below shows the consistency of parameter estimates for factors across all statistically significant models.[5] Notice that, of the four factors stepwise found, only one exhibits above average consistency.



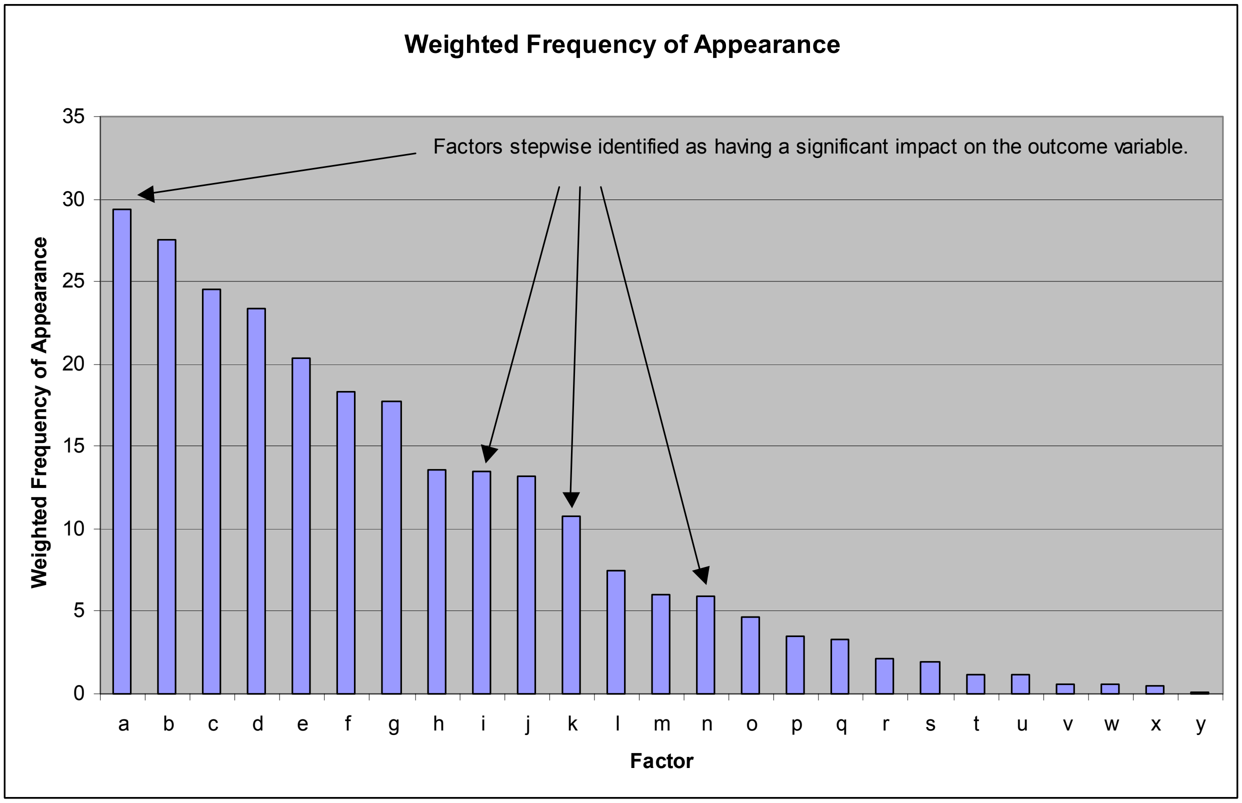
5 For a factor, consistency is measured as the mean of coefficient estimates across models divided by the standard deviation of the coefficient estimates across all models in which the factor appears.

Having all statistically significant models, we can look at the number of times (frequency) a factor appears in a model. The frequency of all factors is shown in the chart below. Notice that only two of the factors stepwise identified appear with above average frequency.



Frequency ignores the impact of a factor on models in which the factor appears. For example, a single factor may appear in many models but always in models with low goodness-of-fit. Conversely, a factor may appear in only a few models yet those models may have very high goodness-of-fit. For this reason, simply counting the number of models in which factors appear can be misleading.

The chart below shows a heuristic that is a "weighted" frequency based on goodness-of-fit.[6] Notice that the factors stepwise identified score worse in terms of weighted frequency than they do in terms of raw frequency.



[6] The weight used is $w = \left|\ln\left(1-\bar{R}^2\right)\right|$.

Conclusion

Both Exhaustive and stepwise regression techniques search for the best possible fit between a dependent variable and a set of potential explanatory variables (factors). While stepwise regression attempts to smartly search a small subset of possible solutions, Exhaustive Regression searches all possible solutions. For a small number of factors, it is more likely that stepwise and Exhaustive Regression will find the same best solution. As the number of factors increases, however, it becomes less probable that stepwise will find the best solution. Regardless of the number of factors, stepwise returns a single result whereas Exhaustive Regression returns all statistically significant results. With all significant results, the analyst can compare the performance of factors across all models – something that is impossible with stepwise. Comparison of results across models can help to identify spurious and unstable results.

	Stepwise Regression	Exhaustive Regression
Examines Every Possible Linear Model		X
Always Finds the Best Fitting Linear Model		X
Returns All Statistically Significant Models		X
Allows Cross-Model Comparisons of All Significant Results		X
Draws Computational Power from a Single Computer	X	
Draws Computational Power from Many Computers Simultaneously		X

The superiority of Exhaustive Regression becomes more pronounced as the number of factors in a data set increases.

A Note on Statistical Searches

All statistical searches (including stepwise regression and Exhaustive Regression) can return results that are simply the result of chance. Such spurious results are of no use. When there is no underlying theoretical model, statistical searches can be used to help the researcher focus further research. Further research is always advisable so as to provide confirmation for results found through a statistical search.