Stepwise vs. Exhaustive Regression

A Comparison using Rates of Return on Bank Stocks

Antony Davies, PhD
June 1, 2001

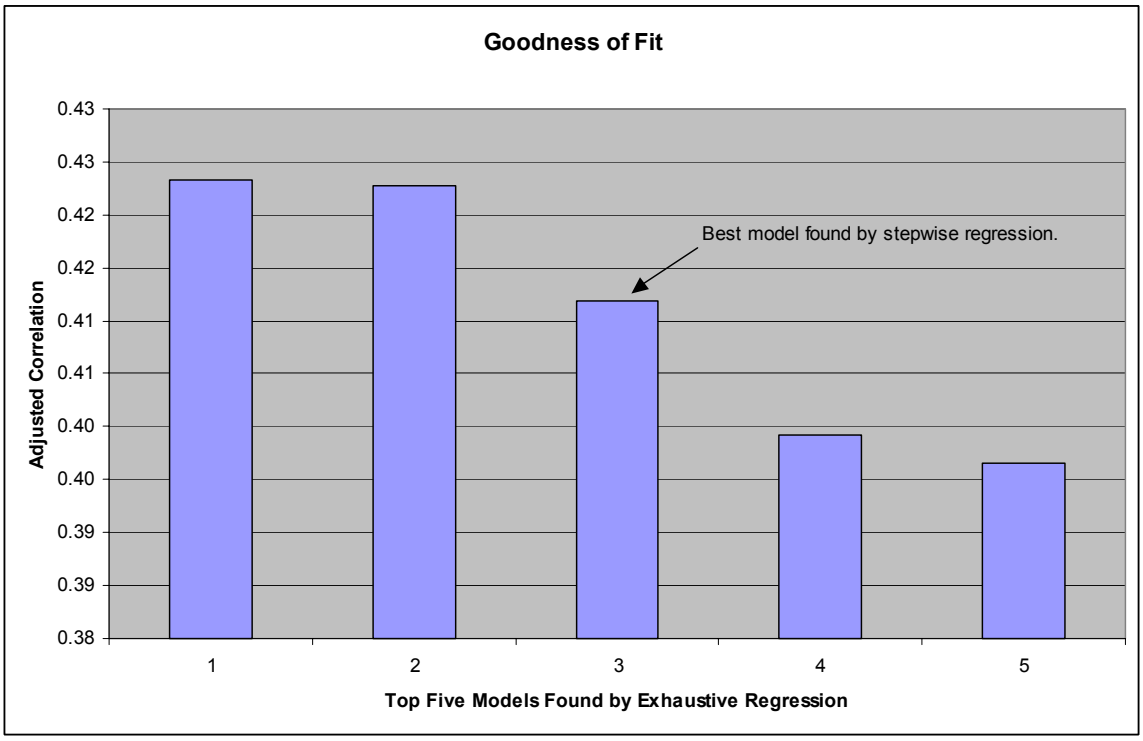
This report compares the results of two analyses of identical stock data using 36 potential explanatory factors from 76 banks in an attempt to find the model that best describes the rates of return on the banks' stocks as a function of the available data. Without recourse to an underlying theoretical 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.

Whereas stepwise techniques use correlations between factors as a means of smartly searching for best-fit models, Exhaustive Regression searches the entire space of potential models and returns those for which all parameter estimates are statistically significant.

With 36 potential factors, there are 69 billion possible linear models. Exhaustive Regression examined all 69 billion models. The analysis, which would have required three and a half years to complete on a single computer, finished in 32 hours. Out of 69 billion possible models, Exhaustive Regression found 152 for which all parameter estimates were significant (significance was measured at the $a = 0.05$ level).

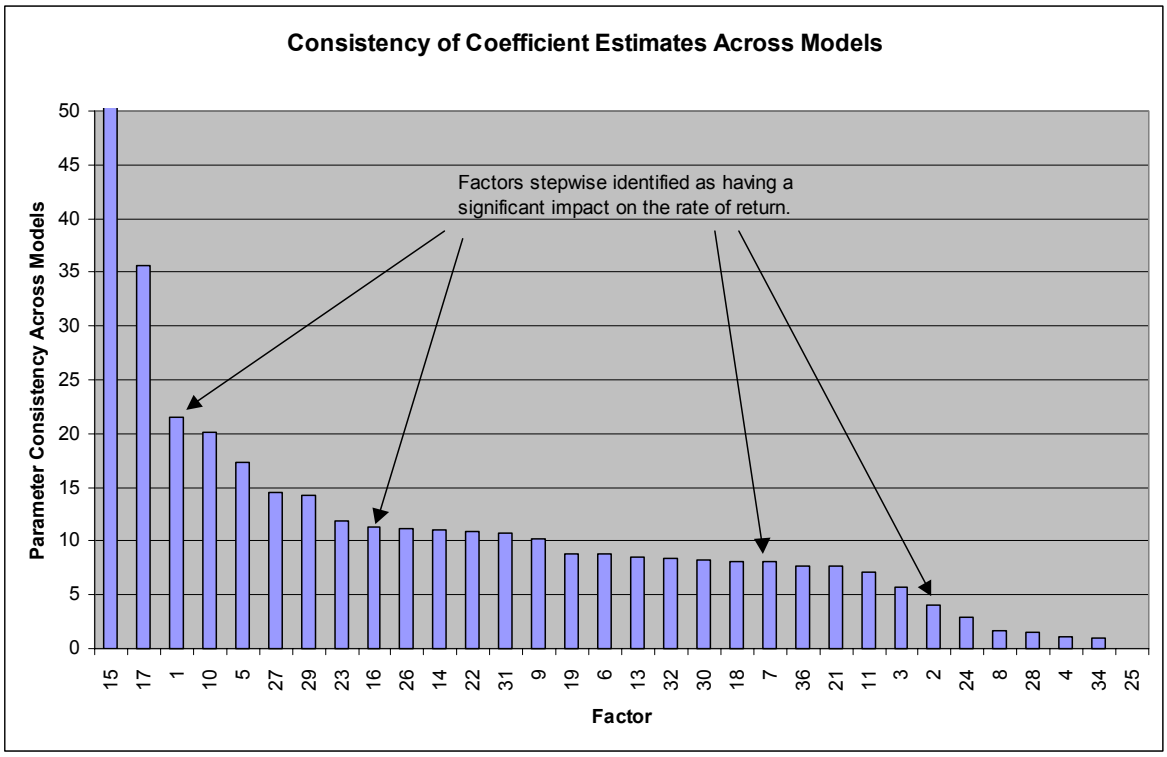
Result #1: Exhaustive Regression found models superior to that found by Stepwise

Of the 152 significant models, two 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 five models found by Exhaustive Regression.



Result #2: Exhaustive Regression found factors that are more stable than those found by Stepwise

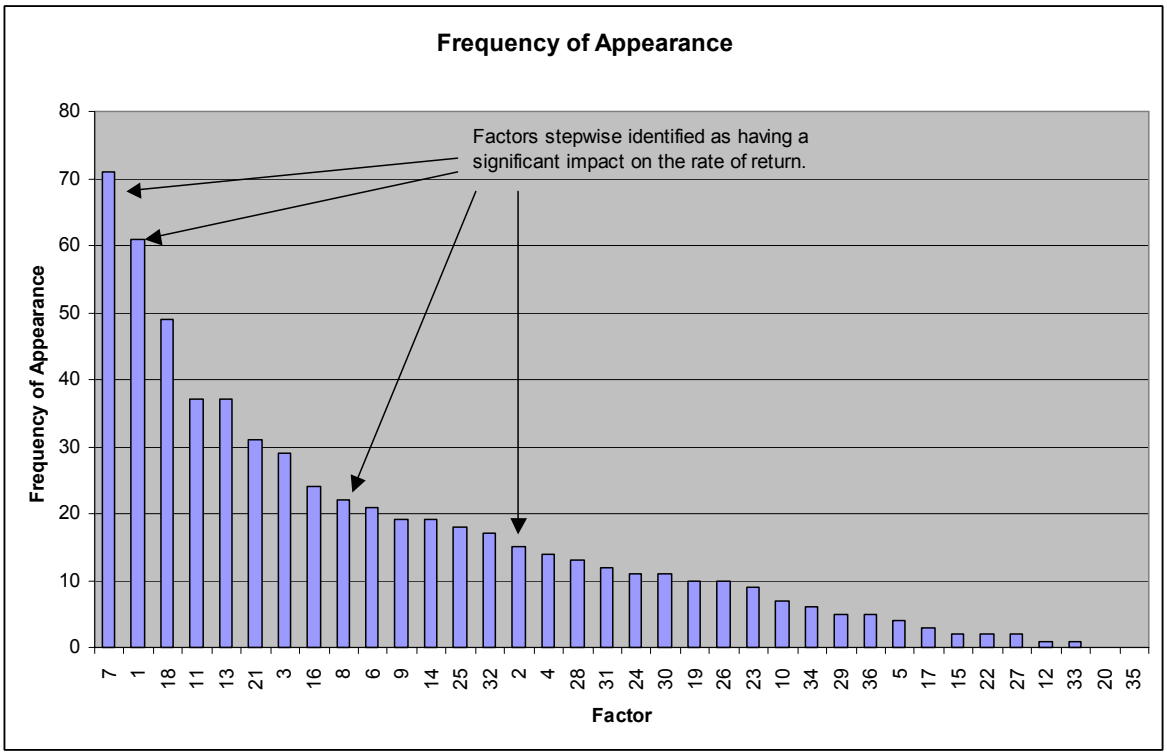
Beyond the ability simply to find better fitting models, Exhaustive Regression allows comparison of the parameter estimates across models – such comparison is impossible with stepwise because stepwise searches only a very small portion of the space of possible solutions. Factors that truly impact the 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.[1] Notice that, of the four factors stepwise found, only two exhibit average or above average consistency.



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 two of the factors stepwise identified appear with more frequency than any other factor. The other two factors identified by stepwise, however, appear with average or below average frequency.

[1] 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.

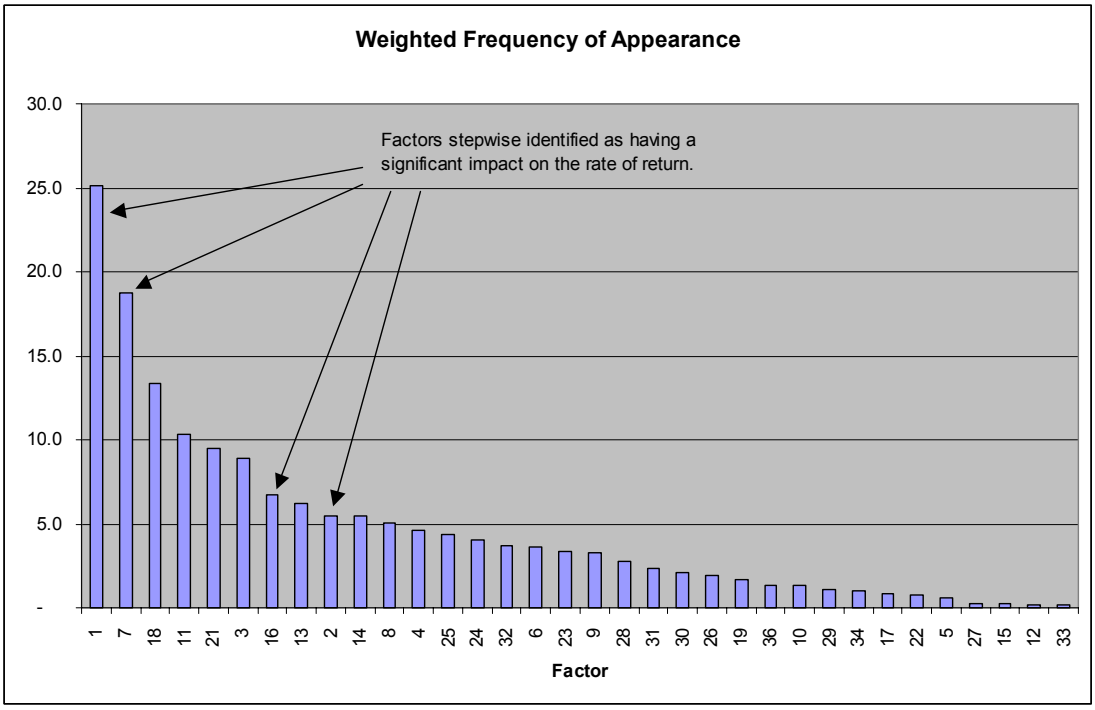
Result #3: Exhaustive Regression found factors that appear in more models than those found by Stepwise



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.

Result #4: Exhaustive Regression found factors with greater weighted frequencies than those found by Stepwise

The chart below shows a heuristic that is a "weighted" frequency based on goodness-of-fit.[2] Notice that the factors stepwise found rate higher on the weighted frequency, yet still are surpassed by factors that stepwise overlooked.



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.

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

	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. For example, in a recent analysis using 70 factors, Exhaustive Regression found 38 models that were superior to the single model found by stepwise regression.

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.