# Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use

Peter L. Flom, National Development and Research Institutes, New York, NY
David L. Cassell, Design Pathways, Corvallis, OR

## ABSTRACT

A common problem in regression analysis is that of variable selection. Often, you have a large number of potential independent variables, and wish to select among them, perhaps to create a 'best' model. One common method of dealing with this problem is some form of automated procedure, such as forward, backward, or stepwise selection. We show that these methods are not to be recommended, and present better alternatives using PROC GLMSELECT and other methods.

**Keywords:** Regression selection forward backward stepwise glmselect.

## INTRODUCTION

In this paper, we discuss variable selection methods for multiple linear regression with a single dependent variable $y$ and a set of independent variables $X_{1:p}$, according to

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

In particular, we discuss various stepwise methods (defined below) and with better alternatives as implemented in SAS®, version 9.1 and PROC GLMSELECT. Stepwise methods are also problematic for other types of regression, but we do not discuss these.

The essential problems with stepwise methods have been admirably summarized by Frank Harrell in Regression Modeling Strategies Harrell (2001), and can be paraphrased as follows:

1. $R^2$ values are biased high
2. The $F$ and $\chi^2$ test statistics do not have the claimed distribution.
3. The standard errors of the parameter estimates are too small.
4. Consequently, the confidence intervals around the parameter estimates are too narrow.
5. p-values are too low, due to multiple comparisons, and are difficult to correct.
6. Parameter estimates are biased high in absolute value.
7. Collinearity problems are exacerbated

This means that your parameter estimates are likely to be too far away from zero; your variance estimates for those parameter estimates are not correct either; so confidence intervals and hypothesis tests will be wrong; and there are no reasonable ways of correcting these problems.

Most devastatingly, it allows the analyst not to think. Put in another way, for a data analyst to use stepwise methods is equivalent to telling his or her boss that his or her salary should be cut. One additional problem is that the methods may not identify sets of variables that fit well, even when such sets exist Miller (2002).

We detail why these methods are poor, and suggest some better alternatives. In the remainder of this section, we discuss the SAS implementation of the stepwise methods. Next, we show how these methods violate statistical theory; then we show that the theoretical violations have important practical consequences in commonly encountered situations. In the penultimate section we briefly discuss some better alternatives, including two that are newly implemented in SAS in PROC GLMSELECT. We close by summarizing our results, making recommendations, and suggesting further readings.

### TERMINOLOGY

A *variable selection method* is a way of selecting a particular set of independent variables (IVs) for use in a regression model. This selection might be an attempt to find a 'best' model, or it might be an attempt to limit the number of IVs when there are too many potential IVs. There are a number of commonly used methods which we call *stepwise techniques*. These include

- Forward selection begins with no variables selected (the null model). In the first step, it adds the most significant variable. At each subsequent step, it adds the most significant variable of those not in the model, until there are no variables that meet the criterion set by the user.
- Backward selection begins with all the variables selected, and removes the least significant one at each step, until none meet the criterion.
- Stepwise selection alternates between forward and backward, bringing in and removing variables that meet the criteria for entry or removal, until a stable set of variables is attained.
- Bivariate screening starts by looking at all bivariate relationships with the DV, and includes any that are significant in a main model.

**SAS IMPLEMENTATION**

SAS implements forward, backward, and stepwise selection in PROC REG with the SELECTION option on the MODEL statement. Default criteria are p = 0.5 for forward selection, p = 0.1 for backward selection, and both of these for stepwise selection. The criteria can be adjusted with the SLENTRY and SLSTAY options.

## WHY THESE METHODS DON'T WORK: THEORY

The essential problem is that we are applying methods intended for one test to many tests. The F-test and all the other statistics generated by PROC GLM or PROC REG are based on a single hypothesis being tested. In stepwise regression, this assumption is grossly violated in ways that are difficult to determine. For example, if you toss a coin ten times and get ten heads, then you are pretty sure that something weird is going on. You can quantify exactly how unlikely such an event is, given that the probability of heads on any one toss is 0.5. If you have 10 people each toss a coin ten times, and *one* of them gets 10 heads, you are less suspicious, but you can still quantify the likelihood. But if you have a bunch of friends (you don't count them) toss coins some number of times (they don't tell you how many) and someone gets 10 heads in a row, you don't even know how suspicious to be. That's stepwise.

As a result of the violation of the assumption, the following can be shown to be true Harrell (2001):

- Standard errors are biased toward 0
- p-values also biased toward 0
- Parameter estimates biased away from 0
- Models too complex

## THESE METHODS REALLY DON'T WORK: EXAMPLES

One test of a technique is whether it works when all the assumptions are precisely met. We generate multivariate data for a that meets all the assumptions of linear regression

1. $\varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$
2. Linearity of relationship between IVs and DV

. In all models, the IVs will be $\overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

For our first example, we ran a regression with 100 subjects and 50 independent variables — all white noise. We used the defaults in stepwise, which are a entry level and stay level of 0.15; in forward, an entry level of 0.50, and in backward a stay level of 0.10. The final stepwise model included 15 IVs, 5 of which were significant at $p < .05$. Forward selection yielded a final model with 29 IVs, 5 sig at $p < .05$. Backward selection yielded 10 IVs, 8 sig at $p < .05$.

Of course, this violates the rule of thumb about how many subjects you should have for each IV. Therefore, for our second example we ran a similar test with 1000 subjects. Results were not encouraging: Stepwise led to 10 IVs with 5 significant at 0.05; forward to 28 IVs, with 5 significant at 0.05, and backward to 10 IVs, with 8 significant at 0.05. This is more or less what we would expect with those p values, but it does not give one much confidence in these methods' abilities to detect signal and noise.

Usually, when one does a regression, at least one of the independent variables is really related to the dependent variable, but there are others that are not related. For our third example we added one real relationship to the above models. However, since measurements contain noise, we also added noise to the model, so that the correlation of the 'real' IV with the DV was 0.32. In this case, with 100 subjects, 50 'false' IVs, and one 'real' one, stepwise selection did not select the real one, but did select 14 false ones. Forward and backward both included the real variable, but forward also included 23 others. Backward did better, including only one false IV. When the number of subjects was increased to 1000, all methods included the real variable, but all also included large numbers of false ones.

That's what happens when the assumptions aren't violated. But sometimes there are problems. For our fourth example we added one outlier, to the example with 100 subjects, 50 false IVs and 1 real IV, the real IV was included, but the parameter estimate for that variable, which ought to have been 1, was 0.72. With two outliers (example 5), the parameter estimate was reduced to 0.44.

## ALTERNATIVES TO STEPWISE METHODS

In this section we review some of the many alternatives to stepwise selection. First, we discuss methods that are not automatic, but that rely on judgement. Then we discuss some automatic methods. However, it is our view that no method can be sensibly applied in a truly automatic manner. The methods we discuss below perform better than stepwise, but their use is not a substitute for substantive and statistical knowledge. A difficulty with evaluating different statistical methods of solving a problem (such as variable selection) is that, to be general, the evaluation should not rely on the particular issues related to a particular problem. However, in actually solving data analytic problems, these particularities are essential.

### A FULL(ER) MODEL

One option that seems to often be neglected in research is leaving non-significant variables in the model. This is problematic in some cases, for example, if there are too many potential IVs, or if the IVs are collinear. However, there is nothing intrinsic in multiple regression that requires only significant IVs to be included. In fact, these IVs may be interesting despite their non-significance. First, including them may affect the parameters of other IVs. Second, if theory suggests that they will be significant (and theory ought to at least suggest this - or why are you including them in the list of potential IVs?) then a small and non-significant result is interesting. Third, although a full discussion of p-values is beyond the scope of this paper, in general it is the size of the parameter estimates that ought to be of most interest, rather than their statistical significance.

The problem with this method is that adding variables to the regression equation increases the variance of the predicted values (see e.g. Miller (2002)) — this is the price paid for the decreased bias in the predicted values. This bias-variance tradeoff is central to the selection of a good method and a good model.

### EXPERT KNOWLEDGE

Another excellent alternative that is often overlooked is using substantive knowledge to guide variable selection. Many researchers seem to believe that the statistical analysis should guide the research; this is rarely the case: Expert knowledge should guide the research. Indeed, this method ought not really be considered an alternative, but almost a prerequisite to good modeling. Although the amount of substantive theory varies by field, even the fields with the least theory must have some, or there would be no way to select variables, however tentatively.

### MODEL AVERAGING

Space does not permit a full discussion of model averaging here, but the central idea is to first develop a set of plausible models, specified independently of the sample data, and then obtain a plausibility index for each model. This index can, for example, be based on Akaike Information Criterion weights given by

$$w_i = \frac{exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R} exp(-\frac{1}{2}\Delta_r)}$$

where $\Delta_r$ are differences in ordered AIC, and R is the number of models to be averaged. Then, these models are combined using:

$$\hat{\bar{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i$$

where $\hat{\theta}_i$ are the parameter estimates from the individual models.

For details, see Burnham & Anderson (2002).

### PARTIAL LEAST SQUARES

When one has too many variables, a standard data reduction technique is principal components analysis (PCA), and some have recommended PCA regression. This involves reducing the number of IVs by using the largest eigenvalues of $\mathbf{X}'\mathbf{X}$. There are two problems with this approach.

- The principal components may have no sensible interpretation
- The dependent variable may not be well predicted by the principal components, even though it would be well predicted by some other linear combination of the independent variables Miller (2002).

Partial least squares finds linear combinations of the IVs that are related to the DV. One way of looking at this is to note that principal component regression is based on the spectral decomposition of $\mathbf{X}'\mathbf{X}$, partial least squares is based on the decomposition of $\mathbf{X}'\mathbf{Y}'$.

### LASSO

The lasso is one of a class of shrinkage methods (perhaps the best-known shrinkage method is ridge regression); the lasso parameter estimates are given Trevor Hastie & Friedman (2001) by

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \beta)0 - \sum_{j=1}^{p} x_{ij}\beta_j^2$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq s$$

where

$N$ is sample size

$y_i$ are values of the dependent variable

$\beta_0$ is a constant, often parameterized to 0 by standardizing the predictors

$x_{ij}$ are the values of the predictor variables

$s$ is a shrinkage factor

### LAR

Least Angle Regression was developed by Efron, Hastie, Johnstone & Tibshirani (2004). It begins by centering all the variables and scaling the covariates. Initially, all parameters are set to 0, and then parameters are added based on correlations with current residual

### CROSS VALIDATION

Cross-validation is a resampling method, like the bootstrap or the jackknife, which takes yet another approach to model evaluation. When people talk about using hold-out samples, this is not really cross-validation. Cross-validation typically takes K replicate samples of the data, each one using (K-1)/K of the data to build the model and the remaining 1/K of the data to test the model in some way. This is called a K-fold cross-validation. For a sample of size N, leave-one-out-cross-validation, or LOOCV, acts a little like a jackknife structure, taking N-1 of the data points to build the model and testing the results against the remaining single data point, in N systematic replicates, with the kth point being dropped in the kth replicate. So this is the K-fold cross-validation taken to its extreme, with K=N. In addition, the random K-fold cross-validation does not split the data into a partition of K subsets, but takes K independent samples of size N*(K-1)/K instead.

## PROC GLMSELECT

The GLMSELECT procedure is experimental in version 9, and must be downloaded from the SAS web site (`www.sas.com`) which also has the documentation. GLMSELECT has many features, and we will not discuss all of them; rather, we concentrate on the three that correspond to the methods just discussed.

The GLMSELECT statement is as follows:

```
PROC GLMSELECT <options>;
CLASS variable;
MODEL variable = <effects></options>;
SCORE <DATA = dataset> <OUT = dataset>;
```

Key options on the GLMSELECT statement include:

- DATA =
- TESTDATA =
- VALDATA =
- PLOTS =

The MODEL statement allows you to choose selection options including:

- Forward
- Backward
- Stepwise
- Lasso
- LAR

and also allows you to select choose options:

- The CHOOSE = criterion option chooses from a list of models based on a criterion
- Available criteria are: adjrsq, aic, aicc, bic, cp ,cv, press, sbc, validate
- CV is residual sum squares based on k-fold CV
- VALIDATE is avg. sq. error for validation data

The STOP criterion option stops the selection process. Available criteria are: adjrsq, aic aicc, bic, cp cv, press, sbc, sl, validate

There are a number of other options to discuss.

- HIERARCHY =
- CVDETAILS= AND CVMETHOD=
- STATS =
- STB

You can combine the options in lots of ways. e.g. `selection = forward(stop = AIC sle = .2)` `selection = forward(st`

When applied to the above problems, with the default options LASSO and LAR performed quite well. We show results for LASSO, results for LAR were almost identical.

- N = 100, 50 IVs, all noise . . . none selected
- N = 1000, 50 IVs, all noise . . . none selected
- N = 100, 50 noise variables, 1 real . . . none selected
- N = 1000, 50 noise variables, 1 real . . . only real selected
- N = 100, 50 noise variables, 1 real, 1 outlier . . . .... param est now .99
- N = 100, 50 noise variables, 1 real, 2 outliers . . . ....no variables included

LAR performed almost identically.

## LIMITATIONS

Although we firmly believe that the Lasso and LAR methods are superior alternatives to other methods, they are not panaceas. The methods still make assumptions, and these assumptions need to be checked. In addition to the standard statistical assumptions, they assume that the models being considered make substantive sense. As Weisberg notes in his discussion of Bradley Efron & Tibshirani (2004), neither LARS nor any other method of automatic method 'has any hope of solving [the problem of model building] because automatic methods by their very nature do not consider the context of the problem at hand'. Or, as the first author put it on SAS-L: 'Solving statistical problems without context is like boxing while blindfolded. You might hit your opponent in the nose, or you might break your hand on the ring post'.

## SUMMARY, RECOMMENDATIONS, AND FURTHER READING

Although no method can substitute for substantive and statistical expertise, LASSO and LAR offer much better alternatives than stepwise as a starting point for further analysis. The wide range of options available in both these methods allows for considerable exploration, and for eliminating models that do not make substantive sense.

For additional information on the problems posed by stepwise, Harrell (2001) offers a relatively nontechnical introduction, together with good advice on regression modeling in general. Burnham & Anderson (2002) offers a more detailed approach; the first chapter outlines the problem, and the remaining chapters offer two general frameworks for solving it (one based on information criteria and the other on multimodel averaging). For more on LASSO and LAR, two key references are **?** and Trevor Hastie & Friedman (2001).

### REFERENCES

Bradley Efron, Trevor Hastie, I. J. & Tibshirani, R. (2004), 'Least angle regression', *Annals of Statistics* **32**, 407–499.

Burnham, K. P. & Anderson, D. R. (2002), *Model selection and multimodel inference*, Springer, New York.

Harrell, F. E. (2001), *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*, Springer-Verlag, New York.

Miller, A. J. (2002), *Subset selection in regression*, Chapman & Hall, London.

Trevor Hastie, R. T. & Friedman, J. (2001), *The elements of statistical learning*, Springer-Verlag, New York.

## CONTACT INFORMATION

Peter L. Flom
515 West End Ave
Apt 8C
New York, NY 10024 peterflomconsulting@mindspring.com
(917) 488 7176

David L. Cassell
mathematical statistician
Design Pathways
3115 NW Norwood Pl.
Corvallis OR 97330