The Stata Journal (2007)
7, Number 2, pp. 272–274

Stata tip 46: Step we gaily, on we go

Richard Williams
University of Notre Dame
Notre Dame, IN 46556
richard.a.williams.5@nd.edu

The `nestreg` and `stepwise` prefix commands allow users to estimate sequences of nested models. With `nestreg`, you specify the order in which variables are added to the model. So, for example, a first model might include only demographic characteristics of subjects, a second could add attitudinal measures, and a third could add interaction terms. Conversely, with `stepwise`, the order in which variables enter the model is determined empirically. With forward selection, the variable or block of variables that most improves fit will be entered first, followed by the variable or variables that most improve fit given the variables already in the model, and so forth. Variables that do not meet some specified level of significance will never enter the model.

Despite their similarities, the two commands differ dramatically in the amount of detail that they provide. `stepwise` gives the estimates for the final model it fits but tells little about the intermediate models other than the order in which variables were entered. `nestreg`, on the other hand, offers a wealth of information. The results from each intermediate model can be printed and their estimates stored for later use. Particularly useful is that `nestreg` offers several measures of contribution to model fit. Wald statistics, likelihood-ratio chi-squareds, R^2 and change-in-R^2 statistics, and Bayesian information criterion (BIC) and Akaike information criterion (AIC) measures are available for each intermediate model. Such measures provide a variety of ways of assessing the importance and effect of each variable or set of variables added to the model.

When forward selection is used, there is a relatively easy way to make the results from `stepwise` as informative and detailed as those provided by `nestreg`. Simply fit the models with `stepwise`, and then refit the models with `nestreg`, listing variables in the order they were added by `stepwise`. For example,

© 2007 StataCorp LP

```
. sysuse auto
(1978 Automobile Data)
. stepwise, pe(.05): regress price mpg weight length foreign
                     begin with empty model
p = 0.0000 <  0.0500  adding  weight
p = 0.0000 <  0.0500  adding  foreign
p = 0.0069 <  0.0500  adding  length
```

Source	SS	df	MS		
Model	348565467	3	116188489	Number of obs =	74
Residual	286499930	70	4092856.14	F(3, 70) =	28.39
				Prob > F =	0.0000
Total	635065396	73	8699525.97	R-squared =	0.5489
				Adj R-squared =	0.5295
				Root MSE =	2023.1

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	5.774712	.9594168	6.02	0.000	3.861215	7.688208
foreign	3573.092	639.328	5.59	0.000	2297.992	4848.191
length	-91.37083	32.82833	-2.78	0.007	-156.8449	-25.89679
_cons	4838.021	3742.01	1.29	0.200	-2625.183	12301.22

```
. nestreg, quietly store(m): regress price weight foreign length if e(sample)
Block  1: weight
Block  2: foreign
Block  3: length
```

Block	F	Block df	Residual df	Pr > F	R2	Change in R2
1	29.42	1	72	0.0000	0.2901	
2	29.59	1	71	0.0000	0.4989	0.2088
3	7.75	1	70	0.0069	0.5489	0.0499

Specifying the `quietly` option omitted the estimates from the intermediate models from the output while still showing the various model change statistics. You may wish to omit `quietly` if, for example, changes in coefficients across models as additional variables are added are of interest. The `store(m)` option stored the estimates from the three models as `m1`, `m2`, and `m3`. Storing the results can be useful if we want to replay the results, format the output with some other program, or do more comparisons across models, say, model `m1` versus model `m3`. Using `if e(sample)` guarantees that the same observations are being analyzed by both `nestreg` and `stepwise`. With `stepwise`, observations with missing data on any of the variables specified get excluded from the analysis, even if those variables do not enter the final model. Most critically, the block residual statistics reported by `nestreg` show us how much the addition of each variable increased R^2 and the statistical significance of that change. This approach provides a much more tangible feel for the importance and contribution of each variable than does `stepwise` alone.

If we would also like to see likelihood-ratio contrasts between models, as well as the BIC and AIC statistics for each intermediate model, just add the `lr` option:

```
. nestreg, quietly lr: regress price weight foreign length if e(sample)
Block  1: weight
Block  2: foreign
Block  3: length
```

Block	LL	LR	df	Pr > LR	AIC	BIC
1	-683.0354	25.35	1	0.0000	1370.071	1374.679
2	-670.1448	25.78	1	0.0000	1346.29	1353.202
3	-666.2613	7.77	1	0.0053	1340.523	1349.739

Naturally, keep the usual cautions concerning stepwise procedures in mind. For example, because multiple tests are being conducted, the reported p-values are inaccurate. The researcher may therefore wish to use a more stringent significance level for variable entry, e.g., .01, or use a Bonferroni or other adjustment. Chance alone could cause some variables to enter the model, and a different sample might produce a different final model. Forward and backward selection procedures can also result in different final models. But if you are clear that stepwise selection is appropriate and is being conducted correctly, then combining `stepwise` and `nestreg` should be helpful.