AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

The Stata Journal (2010)
10, Number 4, pp. 650–669

Variable selection in linear regression

Charles Lindsey
StataCorp
College Station, TX
clindsey@stata.com

Simon Sheather
Department of Statistics
Texas A&M University
College Station, TX

Abstract. We present a new Stata program, `vselect`, that helps users perform variable selection after performing a linear regression. Options for stepwise methods such as forward selection and backward elimination are provided. The user may specify Mallows's C_p, Akaike's information criterion, Akaike's corrected information criterion, Bayesian information criterion, or R^2 adjusted as the information criterion for the selection. When the user specifies the `best` subset option, the leaps-and-bounds algorithm (Furnival and Wilson, *Technometrics* 16: 499–511) is used to determine the best subsets of each predictor size. All the previously mentioned information criteria are reported for each of these subsets. We also provide options for doing variable selection only on certain predictors (as in [R] **nestreg**) and support for weighted linear regression. All options are demonstrated on real datasets with varying numbers of predictors.

Keywords: st0213, vselect, variable selection, regress, nestreg

1 Theory/motivation

Redundant predictors in a linear regression yield a decrease in the residual sum of squares (RSS) and less-biased predictions at the cost of an increased variance in predictions.

In settings where there are a small number of predictors, the partial F test can be used to determine whether certain groups of predictors should be included in the model. We divide the predictors into two groups. One group, the base group, will be included in our model. The other group, the suspected group, may or may not be included within the model—we are not yet sure. We call the regression model containing all predictors in both groups, base and suspected, the full (FULL) model. The regression model containing only the base predictors is called the reduced (RED) model.

The partial F test has a test statistic

$$F = \frac{\frac{\mathrm{RSS}_{\mathrm{RED}} - \mathrm{RSS}_{\mathrm{FULL}}}{\mathrm{df}_{\mathrm{RED}} - \mathrm{df}_{\mathrm{FULL}}}}{\frac{\mathrm{RSS}_{\mathrm{FULL}}}{\mathrm{df}_{\mathrm{FULL}}}}$$

Under the null hypothesis that the RED model is true (all the predictor coefficients for the suspected group are zero), F has an $F(\mathrm{df}_{\mathrm{RED}} - \mathrm{df}_{\mathrm{FULL}}, \mathrm{df}_{\mathrm{FULL}})$ distribution. Acceptance of the null hypothesis leads us to use the RED model as our regression model. Rejection of the null hypothesis indicates that we should not ignore the predictors in

© 2010 StataCorp LP

the suspected group (at least one of the predictor coefficients is not zero). We can then reperform the test using subsets of the suspected group to determine which predictors to include in the model. The partial F test may be easily performed in Stata via `nestreg` (see [R] **nestreg**).

In this article, we are concerned with those cases in which there are a large number of predictors. When the suspected predictor list grows large, it is not feasible to use the partial F test method to determine the final regression model. A variety of algorithms have been created to deal with this situation. These variable selection algorithms take the specification of the FULL model and output an optimal RED model. The command presented here, `vselect`, performs the stepwise selection algorithms forward selection and backward elimination as well as the best subsets leaps-and-bounds algorithm.

The output of these algorithms and the partial F test is not very meaningful unless FULL is a valid regression model. A regression model is valid if the assumptions for performing its significance tests are met. They can be accessed using residual plots, scale-location plots, etc. Details can be found in Sheather (2009).

We must also note that inference on the models produced by these algorithms is not equivalent to the inference on the same models that the users find independently without consulting the algorithms. Each step of a variable selection algorithm will fit one or more models and then make an inference on the next step using information from these models. So in addition to inferences made using the final model, many preliminary inferences are made during variable selection.

This will affect the significance levels of the final model. The situation is similar to performing multiple comparisons on the factor means after an analysis of variance tells you there is a significant effect. Each of these comparisons should be evaluated at a different significance level than that of the original factor effect.

Cross-validation methods can be used to handle this multiple inference difficulty. These methods generally perform variable selection on subsets of the data and then use an average measure of the results on these subsets to find the final model. They may also split the data into two parts, performing variable selection on one part (train) and using the other (test) for evaluating the resulting model. Details of this method and a general discussion of the multiple inference problem in variable selection are given in Sheather (2009). The variable selection methods that we use here may be applied under certain cross-validation techniques.

The definition of optimal is not uniformly agreed upon. The optimal model is one that optimizes one or more information criteria. There are multiple information criteria and multiple guidelines for the number and type of information criteria that should be met.

(*Continued on next page*)

1.1 Information criteria

An information criterion is a function of a regression model's explanatory power and complexity. The model's explanatory power (goodness of fit) increases the criterion in the desirable direction, while the complexity of the model counterbalances the explanatory power and moves the criterion in the undesirable direction.

We have singled out five relevant criteria for evaluating linear regression models: Mallows's C_p, ${R^2}_{\text{ADJ}}$ (adjusted), Akaike's information criterion (AIC), Akaike's corrected information criterion (AIC_C), and Bayesian information criterion (BIC). We use the definitions of these criteria given in Sheather (2009) and Izenman (2008). Our definitions for BIC and AIC correspond with those given in `estat` (see [R] **estat**).

The R^2 adjusted information criterion is an improvement to the R^2 measure of a model's explanatory power. We abbreviate the RSS_RED notation to simply RSS. The SST notation refers to the total sum of squares.

$$R^2 = 1 - \frac{\text{RSS}}{\text{SST}}$$

A penalty for unnecessary predictors is introduced by a multiplication by $(n-1)/(n-k-1)$ where n is the sample size and k is the number of predictors in the model.

$${R^2}_{\text{ADJ}} = 1 - \frac{n-1}{n-k-1}\frac{\text{RSS}}{\text{SST}}$$

As ${R^2}_{\text{ADJ}}$ increases, the model becomes more desirable.

The next information criterion, AIC (Akaike 1974), works in the opposite way: as the criterion decreases, the model becomes more desirable. The explanatory power of the model is measured by the maximized log likelihood of the predictor coefficients (assuming a normal model) and error variance. The complexity penalization comes from an addition of the number of predictors.

$$\text{AIC} = 2\left\{-\log L\left(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p, \widehat{\sigma}^2 \mid Y\right) + k + 2\right\}$$

After we formulate the regression model in terms of a normal distribution likelihood, we obtain

$$\text{AIC} = n\log\frac{\text{RSS}}{n} + 2k + n + n\log(2\pi)$$

Hurvich and Tsai (1989) developed a bias-corrected version of AIC, called AIC_C. AIC_C is preferred when the sample size is small or the number of predictors is large relative to sample size. Using our simplified version of AIC,

$$\text{AIC}_\text{C} = \text{AIC} + \frac{2(k+2)(k+3)}{n-(k+2)-1}$$

Let $p = k + 1$. As in the previous section, we use RSS_{FULL} to refer to the RSS under the model containing all predictors. Suppose we have m possible predictors, excluding the intercept. In Izenman (2008), the information criterion C_p, or Mallows's C_p, is defined by

$$C_p = (n - m - 1)\frac{\text{RSS}}{\text{RSS}_{\text{FULL}}} - (n - 2p)$$

According to the C_p criterion, good models have $C_p \approx p$. The full model will always satisfy this criterion. Further, as noted in Hocking (1976), models with small values of Mallows's C_p may be preferred, as well. The Mallows's C_p criterion was originally developed in Mallows (1973).

Our final information criterion, BIC, was proposed by Schwarz (1978). Raftery (1995) provides another development and motivation for the criterion. BIC is similar to AIC, but it adjusts the penalty term for complexity based on the sample size.

$$\text{BIC} = -2 \log L\left(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p, \widehat{\sigma}^2 \mid Y\right) + (k + 2) \log n$$

This reduces to

$$\text{BIC} = n \log \frac{\text{RSS}}{n} + k \log n + n + n \log (2\pi)$$

There is controversy over what should be called the best information criterion. According to Sheather (2009), choosing a model based solely on $R^2{}_{\text{ADJ}}$ generally leads to overfitting (having too many predictors). There is also debate over whether AIC or AIC_{C} should be used in preference to BIC. A comparison of page 46 of Simonoff (2003) with page 208 of Hastie, Tibshirani, and Friedman (2001) demonstrates this. Mallows's C_p suffers from similar controversies. Inference using C_p will be asymptotically equivalent to AIC, but both will share different properties than BIC (Izenman 2008).

For each predictor size k, the best model under each of the information criterions for that predictor size k is the model that minimizes RSS. All other terms are constant for the same predictor size. So at each predictor size, we can find the best model of that size by minimizing the RSS. This remarkable result can greatly simplify the variable selection process.

Now that we have defined the relevant information criteria, we will present the variable selection algorithms implemented in `vselect` that use the criteria. We begin with stepwise selection algorithms.

1.2 Stepwise selection

We present two stepwise selection algorithms, forward selection and backward elimination. These algorithms work with only one information criterion, which may be any of the ones defined previously except Mallows's C_p. Technically, Mallows's C_p could be used in stepwise selection, but the decision on which predictors to keep or add to

the model would be more difficult. All the other criteria measures have an intrinsic ordering among their values. The smallest AIC is best, the larger R^2_{ADJ} is preferable, etc. Mallows's C_p suggests a good model when it is close to the number of predictors and the intercept of the model it measures, but as mentioned in Hocking (1976), small values of Mallows's C_p can yield good models as well. Our stepwise selection algorithms make an automated decision on whether to keep a variable in the model or add a variable to the model. Ideally, this would be based on a simple ranking of the possible models based on an information criterion. If we use both suggestions for interpretation of Mallows's C_p, the algorithm cannot make the decision based on a simple ranking of models. Given this, we will not use Mallows's C_p in stepwise selection. It will still be used in the leaps-and-bounds variable selection, however.

Forward selection is an iterative procedure. Our initial model is composed of only the intercept term. At every iteration, we add to the model the predictor that will yield the most optimal information criterion value when it is included in the model. If there is no predictor that favorably changes the information criterion from its value in the previous iteration, the algorithm terminates with the model from the previous iteration.

Backward elimination is also an iterative procedure. In this case, the initial model is composed of all the predictors. At every iteration, we remove from the model the predictor that will yield the largest improvement in the information criterion value when it is removed from the model. If there is no predictor whose removal will favorably change the information criterion value from that of the previous iteration, the algorithm terminates with the model from the previous iteration.

Both stepwise selection algorithms examine at most $m(m+1)/2$ of the 2^m possible models. When the predictors are highly correlated, the results of stepwise selection and all subsets selection methods can differ dramatically. The algorithms are intuitive and simple to understand. In many cases, they end up with the best model as well.

For a more dependable algorithm, we turn to the leaps-and-bounds algorithm of Furnival and Wilson (1974).

1.3 Leaps and bounds

The leaps-and-bounds algorithm actually gives p different models. Each of the models contains a different number of predictors and is the most optimal model among models having the same number of predictors. The `vselect` command provides the five information criteria for each of the models produced by leaps and bounds. The optimal model is the one model with these qualities: the smallest value of AIC, AIC_{C}, and BIC; the largest value of R^2_{ADJ}; and a value of Mallows's C_p that is close to the number of predictors in the models $+1$ or the smallest among the other Mallows's C_p values. These guidelines help avoid the controversy of which information criterion is the best.

Sometimes there is no single model that optimizes all the criteria. We will see an example of this in the next section. There are no fixed guidelines for this situation. Generally, we can narrow the choices down to a few models that are close in optimization.

Then we make an arbitrary choice among them. All the models in our final group are close together in fit, so we do not lose or gain much explanatory power by choosing one over another.

As explained in Furnival and Wilson (1974), the leaps-and-bounds algorithm organizes all the possible models into tree structures and scans through them, skipping (or leaping) over those that are definitely not optimal. The original description of the algorithm is done with large amounts of Fortran code. Ni and Huo (2005) provide an easier description of the original algorithm.

Each node in the tree corresponds to two sets of predictors. The predictor lists are created based on an automatic ordering of all the predictors by their t test statistic value in the original regression. When the algorithm examines a node, it compares the regressions of each pair of predictor lists with the optimal regressions of each predictor size that have already been conducted. Depending on the results, all or some of the descendants of that node can be skipped by the algorithm. The initial ordering of the predictors and their smart placement in sets within the nodes ensure that the algorithm completes after finding the optimal predictor lists and examining only a fraction of all possible regressions.

Space constraints do not allow us to provide a fuller description of the algorithm than we already have. We can say that it gives us the best models for each predictor quantity and that it does so by only examining a manageable fraction of all the possible models.

1.4 Extensions: Nested models and weighting

Our discussion so far has focused on ordinary least-squares regression models, where variable selection should be performed on all the model predictors. Lawless and Singhal (1978) provides an extension of the leaps-and-bound algorithm to nonnormal models. Rather than using the RSS to compare models, they use the log likelihood $L(\beta)$. An essential condition for our use of the RSS in variable selection is that for a set of predictors A contained in predictor set B, $\text{RSS}(B) \leq \text{RSS}(A)$. In many situations, $L(B) \leq L(A)$, but it is not always true.

Variable selection in weighted linear regressions and in linear regressions where we perform selection on only certain of the predictors will fit into the Lawless and Singhal (1978) theoretical framework and will satisfy the desired likelihood inequality. Weighted linear regression is of tremendous practical use. The form of nested variable selection in which some predictors are fixed is very appealing as well. Through organization or legal policy, analysts may be forced to fix certain predictors as being in their model, but they would still desire to optimize the model with the free predictors to which they have access.

(*Continued on next page*)

`vselect` implements variable selection for weighted linear regression and variable selection where some predictors are fixed. Further implementation of the Lawless and Singhal (1978) methods is under development.

The information criteria will change for weighted linear regression models. Earlier, we simplified the log likelihood of the model in terms of the RSS. Now we will deal with the weighted RSS. Simple derivation will show that our previously presented information criteria formulas are accurate under weighted regression when we substitute weighted RSS for RSS.

We have now explained all the theory behind `vselect`.

2 The vselect command

2.1 Syntax

The syntax for the `vselect` command is

`vselect` *depvar indepvars* [*if*] [*in*] [*weight*] [`, fix(`*varlist*`) best backward forward r2adj aic aicc bic`]

2.2 Options

`fix(`*varlist*`)` fixes these predictors in every regression.

`best` gives the best model for each quantity of predictors.

`backward` selects a model by backward elimination.

`forward` selects a model by forward selection.

`r2adj` uses R^2 adjusted information criterion in stepwise selection.

`aic` uses AIC in stepwise selection.

`aicc` uses AIC_C in stepwise selection.

`bic` uses BIC in stepwise selection.

3 Examples

`vselect` is very straightforward in use. We will first use `bridge.dta` from Sheather (2009) (also Tryfos [1998]). Then we will test `vselect` on two datasets highlighted in Ni and Huo (2005): the diabetes data (Efron et al. 2004) and the famous housing data (Frank and Asuncion 2010). Finally, we will work with a weighted regression from a Stata example dataset that provides state-level information from the 1980 U.S. Census.

3.1 Bridge example

`bridge.dta` can be analyzed using least-squares regression. As Sheather (2009) suggests, we will work with logs of the original predictors.

```
. use bridge
. foreach var of varlist time-spans {
  2. quietly replace `var´ = ln(`var´)
  3. }
. regress time darea-spans
```

Source	SS	df	MS	Number of obs =	45
Model	13.3303983	5	2.66607966	F(5, 39) =	27.05
Residual	3.84360283	39	.098553919	Prob > F =	0.0000
				R-squared =	0.7762
				Adj R-squared =	0.7475
Total	17.1740011	44	.390318208	Root MSE =	.31393

time	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
darea	-.0456443	.1267496	-0.36	0.721	-.3020196	.2107309
ccost	.1960863	.1444465	1.36	0.182	-.0960843	.488257
dwgs	.8587948	.2236177	3.84	0.000	.4064852	1.311104
length	-.0384353	.1548674	-0.25	0.805	-.3516842	.2748135
spans	.23119	.1406819	1.64	0.108	-.0533659	.515746
_cons	2.2859	.6192558	3.69	0.001	1.033337	3.538463

```
. estat vif
```

Variable	VIF	1/VIF
ccost	8.48	0.117876
length	8.01	0.124779
darea	7.16	0.139575
spans	3.88	0.257838
dwgs	3.41	0.293350
Mean VIF	6.19	

Analysis of the residuals and other checks will reveal that the model is valid. As we see, it does have serious multicollinearity problems. All but two of the variance inflation factors exceed 5. Removing redundant predictors should solve this problem.

(*Continued on next page*)

Forward selection

First, we will try to use forward selection based on AIC.

```
. vselect time-spans, forward aic
FORWARD variable selection
Information Criteria: AIC
```

```
Stage 0 reg time  : AIC  86.35751
```

```
AIC   47.19052  :         add      darea
AIC   37.60067  :         add      ccost
AIC   32.80693  :         add       dwgs
AIC   49.00033  :         add     length
AIC   56.43028  :         add      spans
```

```
Stage 1 reg time dwgs : AIC  32.80693
```

```
AIC   30.30586  :         add      darea
AIC   26.61563  :         add      ccost
AIC   28.33827  :         add     length
AIC   25.33412  :         add      spans
```

```
Stage 2 reg time dwgs spans : AIC  25.33412
```

```
AIC   27.12765  :         add      darea
AIC   25.2924   :         add      ccost
AIC   27.14563  :         add     length
```

```
Stage 3 reg time dwgs spans ccost : AIC   25.2924
```

```
AIC   27.06413  :         add      darea
AIC   27.1425   :         add     length
```

```
Final Model
```

Source	SS	df	MS		
Model	13.3047499	3	4.43491664	Number of obs	= 45
Residual	3.86925122	41	.094371981	F(3, 41)	= 46.99
				Prob > F	= 0.0000
Total	17.1740011	44	.390318208	R-squared	= 0.7747
				Adj R-squared	= 0.7582
				Root MSE	= .3072

time	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dwgs	.8355863	.2135074	3.91	0.000	.4043994	1.266773
spans	.1962899	.1107299	1.77	0.084	-.0273336	.4199134
ccost	.148275	.1074829	1.38	0.175	-.0687911	.365341
_cons	2.331693	.3576636	6.52	0.000	1.609377	3.05401

We begin with no predictors, with an AIC of 86.35751 for the intercept in stage 0. Addition of `dwgs` will change the AIC of the model to 32.80693, a more optimal value than the other possibilities of single-predictor addition and the null model. So we add `dwgs` to the model and move to the next stage. When we add `spans` to the model that predicts `time` with `dwgs`, we get an AIC of 25.33412.

So we enter stage 2 with the model predicting `time` by `dwgs` and `spans`. This model yields an AIC of 25.33412. If we add `darea` to this model, we obtain an AIC of 27.12765. Addition of `length` would cause the AIC to rise to 27.14563. Adding either of these would not improve the fit of the model. The addition of the other remaining potential predictor, `ccost`, yields an AIC of 25.2924. This is a very slight gain in terms of AIC, but it is a gain.

In stage 3, we have added `ccost` to the model, so the AIC is now 25.2924. We now predict `spans` based on `dwgs`, `spans`, `ccost`, and the intercept. Addition of `darea` to this model raises the AIC to 27.06413. Addition of `length` to this model raises the AIC to 27.1425. Adding any more predictors causes an increase in AIC, so we terminate the forward selection algorithm with the final model predicting `spans` with `dwgs`, `spans`, and `ccost`.

Now we will compare this result with forward selection using BIC as an information criterion.

```
. vselect time-spans, forward bic
FORWARD variable selection
Information Criteria: BIC
```

```
Stage 0 reg time  : BIC  88.16417
```

```
BIC   50.80385  :         add      darea
BIC   41.21399  :         add      ccost
BIC   36.42026  :         add       dwgs
BIC   52.61365  :         add     length
BIC   60.04361  :         add      spans
```

```
Stage 1 reg time dwgs : BIC  36.42026
```

```
BIC   35.72585  :         add      darea
BIC   32.03562  :         add      ccost
BIC   33.75826  :         add     length
BIC   30.75411  :         add      spans
```

```
Stage 2 reg time dwgs spans : BIC  30.75411
```

```
BIC   34.3543   :         add      darea
BIC   32.51905  :         add      ccost
BIC   34.37228  :         add     length
Final Model
```

Source	SS	df	MS		
Model	13.1251524	2	6.56257622	Number of obs	= 45
Residual	4.0488487	42	.096401159	F(2, 42)	= 68.08
				Prob > F	= 0.0000
				R-squared	= 0.7642
				Adj R-squared	= 0.7530
Total	17.1740011	44	.390318208	Root MSE	= .31049

time	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dwgs	1.041632	.1541992	6.76	0.000	.7304454	1.352819
spans	.2853049	.0909484	3.14	0.003	.1017636	.4688462
_cons	2.661732	.2687132	9.91	0.000	2.119447	3.204017

This method suggests the two-predictor model that predicts `spans` with `dwgs` and `spans`.

Backward elimination

Backward elimination based on AIC yields the same model as forward selection. It takes one fewer iteration.

```
. vselect time-spans, backward aic
BACKWARD variable selection
Information Criteria: AIC
```

```
Stage 0 reg time darea ccost dwgs length spans : AIC  28.99311
```

```
AIC   27.1425   :         remove      darea
AIC   29.07072  :         remove      ccost
AIC   41.42757  :         remove       dwgs
AIC   27.06413  :         remove     length
AIC   30.00605  :         remove      spans
```

```
Stage 1 reg time darea ccost dwgs spans : AIC  27.06413
```

```
AIC   25.2924   :         remove      darea
AIC   27.12765  :         remove      ccost
AIC   39.44412  :         remove       dwgs
AIC   28.60344  :         remove      spans
```

```
Stage 2 reg time ccost dwgs spans : AIC   25.2924
```

```
AIC   25.33412  :         remove      ccost
AIC   37.57602  :         remove       dwgs
AIC   26.61563  :         remove      spans
```

```
Final Model
      Source |       SS       df       MS              Number of obs =      45
-------------+------------------------------           F(  3,    41) =   46.99
       Model |  13.3047499     3  4.43491664           Prob > F      =  0.0000
    Residual |  3.86925122    41  .094371981           R-squared     =  0.7747
-------------+------------------------------           Adj R-squared =  0.7582
       Total |  17.1740011    44  .390318208           Root MSE      =   .3072

------------------------------------------------------------------------------
        time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ccost |    .148275   .1074829     1.38   0.175    -.0687911     .365341
        dwgs |   .8355863   .2135074     3.91   0.000     .4043994    1.266773
       spans |   .1962899   .1107299     1.77   0.084    -.0273336    .4199134
       _cons |   2.331693   .3576636     6.52   0.000     1.609377     3.05401
------------------------------------------------------------------------------
```

In the initial stage, we have the full model with all predictors and an AIC of 28.99311. Removal of `length` will yield the most optimal AIC.

At stage 1, we have removed `length` and our model now has an AIC of 27.06413. If we remove `darea`, we will have reached the final model for forward selection under AIC. Removal of the other predictors will yield less optimal models. At stage 2, removal of any of the predictors will yield worse models in terms of AIC.

Best subsets

The leaps-and-bounds algorithm finds the same forward selection and backward elimination models that we previously discussed. To reach the result, the algorithm needs to perform only 5 out of all 32 possible regressions.

```
. vselect time-spans, best
Response :             time
Fixed Predictors :
Selected Predictors:   dwgs spans ccost darea length

Actual Regressions   5
Possible Regressions 32

 Optimal Models Highlighted:

  # Preds     R2ADJ         C       AIC      AICC       BIC
        1    .70224  9.708371  32.80693  161.0968  36.42026
        2  .7530191  2.082574  25.33412  154.0386  30.75411
        3  .7582178  2.260247   25.2924  154.5353  32.51905
        4  .7534273  4.061594  27.06413  156.9791  36.09744
        5  .7475037         6  28.99311  159.7246  39.83309

Selected Predictors

1  :  dwgs
2  :  dwgs spans
3  :  dwgs spans ccost
4  :  dwgs spans ccost darea
5  :  dwgs spans ccost darea length
```

The optimal $R^2{}_{\text{ADJ}}$ value, 0.7582178, is obtained by the three-variable model with predictors `dwgs`, `spans`, and `ccost`. This is the same model obtained by forward selection and backward elimination under AIC. This model also optimizes AIC, with an AIC of 25.2924.

The most optimal model under BIC and AIC_C is the predictor model using `dwgs` and `spans`. This is the same model found by forward selection under BIC. We find that Mallows's C_p suggests the five-predictor model when we choose the best model as having a C_p value close to the predictor size +1. Otherwise, when picking the smallest Mallows's C_p model, we would choose the two-predictor model that BIC and AIC_C chose.

This is one of the occasions when there is no completely clear, best final model. We can narrow our decision down to the two mentioned models. We might investigate whether AIC_C is more appropriate than AIC in this situation. Recall that picking the model with the highest $R^2{}_{\text{ADJ}}$ generally leads to overfitting (Sheather 2009). Regardless, there is little difference between the values of AIC and $R^2{}_{\text{ADJ}}$ for the two- and three-predictor models. We will arbitrarily pick the two-predictor model that estimates `time` by `dwgs` and `spans` as our final model. This selection yields no high variance inflation factors.

(*Continued on next page*)

```
. estat vif
```

Variable	VIF	1/VIF
dwgs	1.66	0.603451
spans	1.66	0.603451
Mean VIF	1.66	

3.2 Diabetes and housing data

For brevity, we will omit stepwise model selection and focus solely on a best subsets selection method in each of the following datasets. We will document that our implementation of the leaps-and-bounds algorithm obtains the same models as Ni and Huo (2005). We will also demonstrate how few models (relative to all possible models) the leaps-and-bounds algorithm needs to fit before finding the optimal models.

`diabetes.dta` (Efron et al. 2004) contains information on 442 diabetes patients. They are measured on 10 baseline predictor variables and one measure of disease progression. The predictors include age, sex, body mass index (`bmi`), blood pressure (`bp`), and six serum measurements (`s1`–`s6`). The progression variable, `prog`, is our models' response and was recorded a year after the 10 baseline predictors.

Evaluation of the residual plots and other diagnostics does show that the full model is valid. As we see in the variance inflation factors, though, there are serious multicollinearity problems.

```
. use diabetes, clear
. regress prog age-s6
```

Source	SS	df	MS	Number of obs =	442
				F(10, 431) =	46.27
Model	1357023.32	10	135702.332	Prob > F =	0.0000
Residual	1263985.8	431	2932.68168	R-squared =	0.5177
				Adj R-squared =	0.5066
Total	2621009.12	441	5943.33135	Root MSE =	54.154

prog	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0363613	.2170414	-0.17	0.867	-.4629526	.3902301
sex	-22.85965	5.835821	-3.92	0.000	-34.32986	-11.38944
bmi	5.602962	.7171055	7.81	0.000	4.193503	7.012421
bp	1.116808	.2252382	4.96	0.000	.6741061	1.55951
s1	-1.089996	.5733318	-1.90	0.058	-2.21687	.0368782
s2	.7464501	.5308344	1.41	0.160	-.296896	1.789796
s3	.3720042	.7824638	0.48	0.635	-1.165915	1.909924
s4	6.533831	5.958638	1.10	0.273	-5.177772	18.24543
s5	68.48312	15.66972	4.37	0.000	37.68454	99.28169
s6	.2801171	.273314	1.02	0.306	-.257077	.8173111
_cons	-334.5671	67.45462	-4.96	0.000	-467.148	-201.9862

```
. estat vif
    Variable |       VIF       1/VIF
-------------+----------------------
          s1 |     59.20    0.016891
          s2 |     39.19    0.025515
          s3 |     15.40    0.064926
          s5 |     10.08    0.099246
          s4 |      8.89    0.112473
         bmi |      1.51    0.662499
          s6 |      1.48    0.673572
          bp |      1.46    0.685200
         sex |      1.28    0.782429
         age |      1.22    0.821486
-------------+----------------------
    Mean VIF |     13.97
```

When we invoke `vselect` on the data, we find that we only needed to run 29 of a possible 1,024 regressions. Our model choices match those of Ni and Huo (2005). The choices of best model predictor sizes were five for BIC, six for AIC and AIC_C, and eight for $R^2{}_{\text{ADJ}}$. Mallows's C_p chooses the 11-predictor model when we choose the best model as having a C_p value close to the predictor size +1. If we go with the smallest Mallows's C_p value, then we choose the six-predictor model. The six-predictor model seems like a prudent choice, given all of this and the closeness of the optimal BIC and $R^2{}_{\text{ADJ}}$ values to their values under six predictors.

```
. vselect prog age-s6, best
Response :             prog
Fixed Predictors :
Selected Predictors:   bmi bp s5 sex s1 s2 s4 s6 s3 age

Actual Regressions   29
Possible Regressions 1024

 Optimal Models Highlighted:

  # Preds     R2ADJ         C       AIC      AICC       BIC
       1   .3424327  148.3513  4912.038  6166.435  4920.221
       2   .4570228  47.07119  4828.398  6082.832  4840.672
       3   .4765213  30.66302  4813.226  6067.705  4829.591
       4    .487366  21.99793  4804.963  6059.498  4825.419
       5   .5029966  9.147958  4792.264  6046.863  4816.811
       6   .5081925  5.560187  4788.603  6043.278  4817.243
       7   .5084884  6.303253   4789.32  6044.079  4822.051
       8   .5085553  7.248507  4790.241  6045.093  4827.062
       9   .5076694  9.028067  4792.015   6046.97  4832.928
      10   .5065593        11  4793.986  6049.055   4838.99

Selected Predictors

1  :  bmi
2  :  bmi s5
3  :  bmi bp s5
4  :  bmi bp s5 s1
5  :  bmi bp s5 sex s3
6  :  bmi bp s5 sex s1 s2
7  :  bmi bp s5 sex s1 s2 s4
8  :  bmi bp s5 sex s1 s2 s4 s6
9  :  bmi bp s5 sex s1 s2 s4 s6 s3
10 :  bmi bp s5 sex s1 s2 s4 s6 s3 age
```

Using the six-predictor model, we still find some high variance inflation factors between the first and second serum variables. They are far lower in magnitude than they are under the full model:

```
. estat vif
    Variable |       VIF       1/VIF
-------------+----------------------
          s1 |      8.81    0.113561
          s2 |      7.37    0.135750
          s5 |      2.20    0.454745
         bmi |      1.47    0.678813
          bp |      1.34    0.743677
         sex |      1.23    0.815832
-------------+----------------------
    Mean VIF |      3.74
```

If we are concerned about this multicollinearity, we can try the five-predictor model that BIC chose:

```
. estat vif
    Variable |       VIF       1/VIF
-------------+----------------------
          s5 |      1.46    0.684663
          s3 |      1.46    0.685455
         bmi |      1.44    0.692867
          bp |      1.35    0.742260
         sex |      1.24    0.807833
-------------+----------------------
    Mean VIF |      1.39
```

`housing.dta` contains real estate data for 506 Boston residences. You can obtain the dataset at http://archive.ics.uci.edu/ml/datasets/Housing. Many authors have analyzed this dataset (Frank and Asuncion 2010), and we will compare our analysis results with Ni and Huo (2005). Thirteen predictors are used to predict the median value of the home. Using `vselect` on the data, we obtain the same models as Ni and Huo (2005). We performed 71 regressions to obtain the optimal models, which is a small fraction of the total possible number of models that could be fit.

```
. use housing
. vselect y v1-v13, best
Response :             y
Fixed Predictors :
Selected Predictors:   v13 v6 v8 v11 v5 v9 v12 v2 v1 v10 v4 v3 v7
Actual Regressions   71
Possible Regressions 8192
 Optimal Models Highlighted:
  # Preds     R2ADJ         C       AIC      AICC       BIC
        1  .5432418  362.7529  3286.975  4722.989  3295.428
        2  .6371245  185.6474  3171.542  4607.588  3184.222
        3  .6767036  111.6489  3114.097  4550.183  3131.003
        4  .6878351  91.48526  3097.359  4533.493  3118.492
        5  .7051702  59.75364  3069.439  4505.629  3094.798
        6  .7123567  47.17537  3057.939  4494.195  3087.525
        7   .718256  37.05889  3048.438  4484.767  3082.251
        8  .7222072  30.62398  3042.275  4478.685  3080.314
        9  .7252743  25.86591  3037.638  4474.138  3079.903
       10  .7299149  18.20493  3029.997  4466.595  3076.488
       11  .7348058  10.11455  3021.726  4458.432  3072.445
       12  .7343282  12.00275  3023.611  4460.433  3078.556
       13  .7337897        14  3025.609  4462.554   3084.78
Selected Predictors
1  :  v13
2  :  v13 v6
3  :  v13 v6 v11
4  :  v13 v6 v8 v11
5  :  v13 v6 v8 v11 v5
6  :  v13 v6 v8 v11 v5 v4
7  :  v13 v6 v8 v11 v5 v12 v4
8  :  v13 v6 v8 v11 v5 v12 v2 v4
9  :  v13 v6 v8 v11 v5 v9 v12 v1 v4
10 :  v13 v6 v8 v11 v5 v9 v12 v2 v1 v10
11 :  v13 v6 v8 v11 v5 v9 v12 v2 v1 v10 v4
12 :  v13 v6 v8 v11 v5 v9 v12 v2 v1 v10 v4 v3
13 :  v13 v6 v8 v11 v5 v9 v12 v2 v1 v10 v4 v3 v7
```

3.3 Census 1980 Stata dataset

Now we will show how to use the weighting and fixed options for `vselect` by using `census13.dta`, which can be obtained by typing `webuse census13` in Stata or from http://www.stata-press.com/data/r11/census13.dta. This dataset contains one observation per state and records various summary demographic information for the state's population. We wish to predict birthrate `brate` with the median age, `medage`; squared median age, `medage2`; divorce rate, `dvcrate`; marriage rate, `mrgrate`; and geographic region of the state. We standardize median age to prevent obvious multicollinearity between its linear and quadratic term, yielding the transformed variables `tmedage` and `tmedage2`. The 1980 population of the state, `pop`, is used as an analytic weight.

(*Continued on next page*)

```
. webuse census13
(1980 Census data by state)
. describe region
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
region          int    %-8.0g      cenreg     Census region
. label list cenreg
cenreg:
           1 NE
           2 N Cntrl
           3 South
           4 West
. generate ne = region == 1
. generate n = region == 2
. generate s = region == 3
. generate w = region == 4
. summarize medage
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      medage |         50       29.54    1.693445       24.2       34.7
. generate tmedage = (medage-r(mean))/r(sd)
. generate tmedage2 = tmedage^2
```

Invoking `vselect` on the data, we find that AIC and AIC_C both select the five-predictor model. BIC differs in that it chooses to exclude the North Central region of the U.S. as a predictor and so chooses a four-predictor model. $R^2{}_{\text{ADJ}}$ chose to include the marriage rate as a predictor, yielding a six-predictor model. Mallows's C_p advocates the seven-predictor model when we choose a model with C_p close to the number of predictors $+1$. Otherwise, when choosing the smallest C_p value, we will choose the five-predictor model. The level of difference for each criterion from the AIC-chosen predictor size to its own chosen size is minimal. So we choose the five-predictor model. Further investigation will show that this is a valid model. Its variance inflation factors are not problematic, either.

```
. vselect brate tmedage tmedage2 mrgrate dvcrate n s w [aweight=pop], best
Response :              brate
Fixed Predictors :
Selected Predictors:   tmedage tmedage2 n w dvcrate mrgrate s

Actual Regressions   11
Possible Regressions 128

 Optimal Models Highlighted:

  # Preds     R2ADJ         C       AIC      AICC       BIC
        1  .6731149  65.51087    397.97  540.3855   401.794
        2  .7937451  24.89423  375.8925  518.6752  381.6285
        3  .8412783   9.88896  363.7191  506.9766  371.3672
        4  .8557213  6.141906  359.8499  503.6973    369.41
        5  .8623259  5.051247  358.3834  502.9439  369.8555
        6  .8625235  6.012409  359.1621  504.5681  372.5463
        7  .8592919         8  361.1473  507.5412  376.4435

Selected Predictors

1  :  tmedage
2  :  tmedage tmedage2
3  :  tmedage tmedage2 w
4  :  tmedage tmedage2 w dvcrate
5  :  tmedage tmedage2 n w dvcrate
6  :  tmedage tmedage2 n w dvcrate mrgrate
7  :  tmedage tmedage2 n w dvcrate mrgrate s
```

Now suppose that we were forced to include marriage rate as a predictor. We remove it from the predictor list and put it in the `fix()` option.

```
. vselect brate tmedage tmedage2 dvcrate n s w [aweight=pop], best fix(mrgrate)
Response :              brate
Fixed Predictors :     mrgrate
Selected Predictors:   tmedage tmedage2 n w dvcrate s

Actual Regressions   10
Possible Regressions 64

 Optimal Models Highlighted:

  # Preds     R2ADJ         C       AIC      AICC       BIC
        1   .670209  66.15834  399.3598  542.1425  405.0959
        2  .7915307  26.15233  377.3511  520.6086  384.9992
        3  .8385064  11.64741  365.4859  509.3332   375.046
        4  .8565161  6.867985  360.4501  505.0106  371.9222
        5  .8625235  6.012409  359.1621  504.5681  372.5463
        6  .8592919         8  361.1473  507.5412  376.4435

Selected Predictors

1  :  tmedage
2  :  tmedage tmedage2
3  :  tmedage tmedage2 w
4  :  tmedage tmedage2 w dvcrate
5  :  tmedage tmedage2 n w dvcrate
6  :  tmedage tmedage2 n w dvcrate s
```

Here the optimal model on $R^2{}_{\text{ADJ}}$ and AIC and AIC_{C} is the five-predictor model. This is actually a six-predictor model because we have already fixed `mrgrate` as being in the model.

```
. regress brate mrgrate tmedage tmedage2 n w dvcrate [aweight=pop]
(sum of wgt is   2.2591e+08)
      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  6,    43) =   52.24
       Model |  21242.2364     6  3540.37274            Prob > F      =  0.0000
    Residual |   2914.3087    43   67.774621            R-squared     =  0.8794
-------------+------------------------------           Adj R-squared =  0.8625
       Total |  24156.5451    49  492.990717            Root MSE      =  8.2325

------------------------------------------------------------------------------
       brate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mrgrate |  -134.7134   130.6446    -1.03   0.308    -398.1833    128.7566
     tmedage |  -21.11739   1.569742   -13.45   0.000    -24.28307   -17.9517
    tmedage2 |   4.217915   .7312436     5.77   0.000     2.743222    5.692609
           n |    5.03472   2.944985     1.71   0.095    -.9044078    10.97385
           w |   11.92932   3.405185     3.50   0.001     5.062111    18.79653
     dvcrate |   1886.619   735.5317     2.56   0.014     403.2778     3369.96
       _cons |    146.665   4.676581    31.36   0.000     137.2338    156.0962
------------------------------------------------------------------------------
```

4 Conclusion

We explored both the theory and practice of variable selection in linear regression. Using real datasets, we have demonstrated the use of each flavor of variable selection: forward selection, backward elimination, and best subset selection. Variable selection on weighted linear regression and fixed predictor models was also demonstrated.

The `vselect` command was fully defined as a method for performing linear regression variable selection in Stata. Its use on each of the three algorithms and contexts of variable selection was demonstrated using a variety of datasets.

5 References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression. *Annals of Statistics* 32: 407–499.

Frank, A., and A. Asuncion. 2010. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Housing.

Furnival, G. M., and R. W. Wilson. 1974. Regression by leaps and bounds. *Technometrics* 16: 499–511.

Hastie, T., R. J. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hocking, R. R. 1976. A *Biometrics* invited paper: The analysis and selection of variables in linear regression. *Biometrics* 32: 1–49.

Hurvich, C. M., and C.-H. Tsai. 1989. Regression and time series model selection in small samples. *Biometrika* 76: 297–307.

Izenman, A. J. 2008. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. New York: Springer.

Lawless, J. F., and K. Singhal. 1978. Efficient screening of nonnormal regression models. *Biometrics* 34: 318–327.

Mallows, C. L. 1973. Some comments on C_p. *Technometrics* 15: 661–675.

Ni, X., and X. Huo. 2005. Enhanced leaps-and-bounds method in subset selections with additional optimality tests. http://www3.informs.org/site/qsr/downloadfile.php?i=17e62166fc8586dfa4d1bc0e1742c08b.

Raftery, A. E. 1995. Bayesian model selection in social research. *Sociological Methodology* 25: 111–163.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.

Sheather, S. J. 2009. *A Modern Approach to Regression with R*. New York: Springer.

Simonoff, J. S. 2003. *Analyzing Categorical Data*. New York: Springer.

Tryfos, P. 1998. *Methods for Business Analysis and Forecasting: Text and Cases*. New York: Wiley.

About the authors

Charles Lindsey is a statistician and software developer at StataCorp. He graduated from the Department of Statistics at Texas A&M with a PhD in May 2010.

Simon Sheather is professor and head of the Department of Statistics at Texas A&M University. Simon's research interests are in the fields of nonparametric and robust statistics and flexible regression methods. In 2001, Simon was named an honorary fellow of the American Statistical Association. Simon is currently listed on ISIHighlyCited.com among the top one-half of one percent of all mathematical scientists in terms of citations of his published works.