



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.*

THE STATA JOURNAL

Editors

H. JOSEPH NEWTON
Department of Statistics
Texas A&M University
College Station, Texas
editors@stata-journal.com

NICHOLAS J. COX
Department of Geography
Durham University
Durham, UK
editors@stata-journal.com

Associate Editors

CHRISTOPHER F. BAUM, Boston College
NATHANIEL BECK, New York University
RINO BELLOCCO, Karolinska Institutet, Sweden, and
University of Milano-Bicocca, Italy
MAARTEN L. BUIS, University of Konstanz, Germany
A. COLIN CAMERON, University of California–Davis
MARIO A. CLEVES, University of Arkansas for
Medical Sciences
WILLIAM D. DUPONT, Vanderbilt University
PHILIP ENDER, University of California–Los Angeles
DAVID EPSTEIN, Columbia University
ALLAN GREGORY, Queen's University
JAMES HARDIN, University of South Carolina
BEN JANN, University of Bern, Switzerland
STEPHEN JENKINS, London School of Economics and
Political Science
ULRICH KOHLER, University of Potsdam, Germany

FRAUKE KREUTER, Univ. of Maryland–College Park
PETER A. LACHENBRUCH, Oregon State University
JENS LAURITSEN, Odense University Hospital
STANLEY LEMESHOW, Ohio State University
J. SCOTT LONG, Indiana University
ROGER NEWSON, Imperial College, London
AUSTIN NICHOLS, Urban Institute, Washington DC
MARCELLO PAGANO, Harvard School of Public Health
SOPHIA RABE-HESKETH, Univ. of California–Berkeley
J. PATRICK ROYSTON, MRC Clinical Trials Unit,
London
PHILIP RYAN, University of Adelaide
MARK E. SCHAFFER, Heriot-Watt Univ., Edinburgh
JEROEN WEESIE, Utrecht University
IAN WHITE, MRC Biostatistics Unit, Cambridge
NICHOLAS J. G. WINTER, University of Virginia
JEFFREY WOOLDRIDGE, Michigan State University

Stata Press Editorial Manager

LISA GILMORE

Stata Press Copy Editors

DAVID CULWELL, SHELBI SEINER, and DEIRDRE SKAGGS

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go “beyond the Stata manual” in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

<http://www.stata-journal.com>

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

<http://www.stata.com/bookstore/sj.html>

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		Printed & electronic	
1-year subscription	\$115	1-year subscription	\$145
2-year subscription	\$210	2-year subscription	\$270
3-year subscription	\$285	3-year subscription	\$375
1-year student subscription	\$ 85	1-year student subscription	\$115
1-year institutional subscription	\$345	1-year institutional subscription	\$375
2-year institutional subscription	\$625	2-year institutional subscription	\$685
3-year institutional subscription	\$875	3-year institutional subscription	\$965
Electronic only		Electronic only	
1-year subscription	\$ 85	1-year subscription	\$ 85
2-year subscription	\$155	2-year subscription	\$155
3-year subscription	\$215	3-year subscription	\$215
1-year student subscription	\$ 55	1-year student subscription	\$ 55

Back issues of the *Stata Journal* may be ordered online at

<http://www.stata.com/bookstore/sjj.html>

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

<http://www.stata-journal.com/archives.html>

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.



Copyright © 2015 by StataCorp LP

Copyright Statement: The *Stata Journal* and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

The articles appearing in the *Stata Journal* may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the *Stata Journal*, in whole or in part, on publicly accessible websites, file servers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, **STATA**, Stata Press, Mata, **MATA**, and NetCourse are registered trademarks of StataCorp LP.

Best subsets variable selection in nonnormal regression models

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

Simon Sheather
Texas A&M Statistics
College Station, TX

Abstract. We present a new program, `gvselect`, that helps users perform variable selection in regression. Best subsets variable selection is performed and provides the user with the best combinations of predictors for each level of model complexity. The leaps-and-bounds (Furnival and Wilson, 1974, *Technometrics* 16: 499–511) algorithm is applied using the log likelihoods of candidate models. This allows the user to perform variable selection on a wide variety of normal and non-normal regression models. Our method is described in Lawless and Singhal (1978, *Biometrics* 34: 318–327).

Keywords: st0413, `gvselect`, `regress`, `vselect`, variable selection

1 Theory and motivation

Redundant predictors in a regression can yield an increase in the log likelihood and less biased predictions, but they may increase the variance of predictions. In this section, we will discuss the problem of variable selection and decide which predictors to use in a regression. Then, we will introduce a new command for performing variable selection, `gvselect`. We will follow this with examples of `gvselect` using real datasets.

In settings with few predictors, the likelihood-ratio (LR) 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 suspect group”, may or may not be included in the model. We are not yet sure. We call the regression model containing all predictors in both the base and suspected groups “the full model”. The regression model containing only the base predictors is called “the reduced model”.

Let L_f and L_r be the log-likelihood values associated with the full and reduced models, respectively. Let n_f and n_r be the respective number of parameters in the full and reduced models. The LR test statistic is $LR = -2(L_r - L_f)$. Under the null hypothesis that the reduced model is true (all the predictor coefficients for the suspected group are zero), LR has an χ^2 distribution with $n_f - n_r$ degrees of freedom. Accepting the null hypothesis leads us to use the reduced model as our regression model.

Rejecting the null hypothesis indicates that we should not ignore the predictors in 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 LR test may be easily performed in Stata via `lrtest` (see [R] `lrtest`).

This article is about settings with a large number of predictors. When the suspected predictor list grows large, it is not feasible to use the LR test method to determine the best regression model.

1.1 Information criteria

The definition of the best, or the optimal, model will vary with the information criteria used for evaluating models. 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. Information criteria help us determine an optimal tradeoff between prediction accuracy and precision.

Akaike's (1974) information criterion (AIC) is a popular criterion for comparing different models. As AIC 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 parameters p .

$$\text{AIC} = 2(-L_r + p)$$

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

$$\text{BIC} = -2L_r + p \log n$$

AIC and BIC can be estimated within Stata using `estat ic`. There is debate over whether AIC should be used in preference to BIC. A comparison of page 46 of Simonoff (2003) and page 235 of Hastie, Tibshirani, and Friedman (2009) demonstrates this. We find that selection based on BIC yields more parsimonious models here, but that does not mean these models are superior to the more complex models that are selected based on AIC.

The `gvselect` command calculates only AIC and BIC. Those interested in other information criteria and linear regression models should investigate the `vselect` command. This command was introduced in Lindsey and Sheather (2010b). It performs linear regression variable selection using the information criteria AIC, corrected AIC (Hurvich and Tsai 1989), BIC, R^2 adjusted (Sheather 2009), and Mallows's C (Izenman 2008).

1.2 Variable selection algorithms

A variety of algorithms have been created to perform variable selection using information criteria like AIC and BIC. These variable selection algorithms take the specification of the full model and output a reduced model.

Stepwise selection algorithms use backward elimination or forward selection and add or remove predictors iteratively. They may yield a reasonable model but are not guaranteed to select an information criterion's optimal model.

The algorithm implemented in `gvselect` is guaranteed to select the best models for BIC and AIC. At each level of complexity (number of parameters p), the optimal model is the model with the largest log likelihood. `gvselect` applies the leaps-and-bounds algorithm (Furnival and Wilson 1974) to the log likelihoods of the candidate models. The algorithm outputs the model with the largest log likelihood for each level of complexity. This approach is described in Lawless and Singhal (1978). The AIC and BIC of these final models can then be compared. Furnival and Wilson (1974) focus on doing variable selection for linear regression models, and residual sums of squares are used instead of log likelihoods. Lawless and Singhal (1978) generalized their method to nonnormal regression models, using log likelihood instead of residual sum of squares to choose models.

As Furnival and Wilson (1974) explain, the leaps-and-bounds algorithm organizes all the possible models in 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. They use a pair tree where each node has two subsets of predictors.

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 ensures that the algorithm completes after finding the optimal predictor lists and examining only a fraction of all possible regressions.

The predictor lists in the pair tree are created based on an automatic ordering of all the predictors by their χ^2 test statistic value in the original regression. The first node, $(\phi, \{1, \dots, k\})$, contains the empty set and all k predictors. Let (Ω_1, Ω_2) be a node in the pair tree. If (Ω_1, Ω_2) is not the first child of its parent, then it has children. Denote the first child of (Ω_1, Ω_2) as $(\Omega_{11}, \Omega_{21})$. The subset Ω_{11} is obtained by removing the last predictor from Ω_2 . The subset Ω_{21} is obtained by removing the second-to-last predictor from Ω_2 . For $i = 2, \dots, k$, denote child i of (Ω_1, Ω_2) as $(\Omega_{1i}, \Omega_{2i})$. The subset Ω_{1i} is obtained by removing the i th from the last through the last predictor of Ω_2 . So two predictors are removed for a second child, three for a third, etc. The subset Ω_{2i} is obtained by removing the $i + 1$ th from the last predictor of Ω_2 . Figure 1 shows the pair tree for a regression with five predictors.

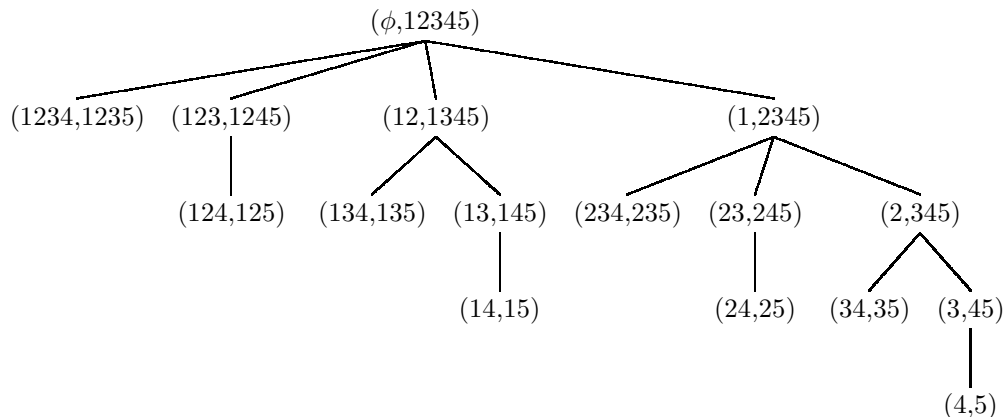


Figure 1. Pair tree for five predictors

We traverse the tree from the root node down and left to right. Let L_i be the maximum log likelihood for i predictors that has so far been calculated in the traversal. We denote the log likelihood under the predictors in Ω_a by L_{Ω_a} . If $\Omega_b \subseteq \Omega_a$, then $L_{\Omega_b} \leq L_{\Omega_a}$. Using this inequality and the pair tree definition, we obtain two rules of traversal. In the first rule, if $L_{\Omega_2} < L_{|\Omega_1|}$, then we can skip all the descendants of (Ω_1, Ω_2) . None of the descendant models have larger likelihoods than those that have already been found. In the second rule, we can skip the first i descendants of (Ω_1, Ω_2) if $L_{|\Omega_2|-i-1} < L_{\Omega_2} \leq L_{|\Omega_2|-i}$. Using these rules, we can find the models with the highest log likelihood for every complexity level and avoid fitting all possible models.

1.3 Model validity and cross-validation

Note that the output of the leaps-and-bounds algorithm and the LR test are not very meaningful unless the full model is a valid regression model. A regression model is valid if the assumptions to perform its significance tests are met. The assumptions for linear regression models can be assessed using residual plots, scale-location plots, etc. Details can be found in Sheather (2009). Diagnostics for logistic regression models are discussed in Hosmer, Lemeshow, and Sturdivant (2013). Diagnostics for categorical models are discussed in Agresti (2013).

Marginal model plots were proposed by Cook and Weisberg (1997). They are a graphical diagnostic tool that can be applied to many regression models. They allow visual comparison of the parametric model fit of the conditional mean with a non-parametric estimate of the conditional mean. The proper specification of the model is corroborated when the two estimates correspond. When they differ, it suggests the model is misspecified. The `mmp` command, developed in Lindsey and Sheather (2010a), can be used to draw marginal model plots in Stata.

We must also note that inference on the models produced by the leaps-and-bounds algorithm is not equivalent to the inference on the same models that the users find independently without consulting the algorithm. 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 for evaluating the resulting model (test). Details of this method and a general discussion of the multiple-inference problem in variable selection are given in Sheather (2009). The variable selection method that we use here may be applied under certain cross-validation techniques.

In the next section, we will provide the full syntax of the `gvselect` command. Then, we will demonstrate how to use `gvselect` with two real-world examples. First, we determine the optimal linear regression model for measuring diabetes progression. Second, we find the best Poisson regression model for predicting the number of doctor visits an individual will have in a two-week period.

2 Use and examples

The `gvselect` command has the following syntax:

```
gvselect <term> varlist [, nmodels(#)] : est_cmd
```

est_cmd is a call to an estimation command that returns a log-likelihood numeric result. *est_cmd* contains instances of `<term>`. These are replaced by variables in *varlist* to determine the best subsets of *varlist* for fitting the model of interest. If the `nmodels()` option is specified with value `#`, the best `#` models are reported for each level of model complexity.

The `gvselect` command is straightforward in use. We will first demonstrate on a dataset highlighted in Ni and Huo (2005), the diabetes data, which were introduced in Efron et al. (2004). Then, we will use `gvselect` on data from Cameron and Trivedi (1986) to perform variable selection for a Poisson regression predicting the number of doctor visits made in a two-week period.

2.1 Diabetes data

The diabetes study data (Efron et al. 2004) contain information on 442 diabetes patients. They are measured on 10 baseline predictor variables and a measure of disease progression. The predictors are 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. We will estimate the parameters of the linear regression of `prog` on the 10 predictors.

Evaluation of the residual plots and other diagnostics do show the full model is valid. But, as we see in the variance inflation factors, there are serious multicollinearity problems. In particular, `s1`–`s5` all exhibit variance inflation factors that exceed 8.

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

Source	SS	df	MS	Number of obs	=	442
Model	1357023.32	10	135702.332	F(10, 431)	=	46.27
Residual	1263985.8	431	2932.68168	Prob > F	=	0.0000
				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	

Now, we invoke `gvselect` on the data. Our model choices match those of Ni and Huo (2005). The choices of best model predictor sizes were 5 for BIC and 6 for AIC. The 6-predictor model seems like a prudent choice, given the closeness of the optimal BIC value to the BIC value under 6 predictors.

```
. gvselect <term> age-s6: regress prog <term>
```

```
Optimal models:
```

# Preds	LL	AIC	BIC
1	-2454.019	4912.038	4920.221
2	-2411.199	4828.398	4840.672
3	-2402.613	4813.226	4829.591
4	-2397.481	4804.963	4825.419
5	-2390.132	4792.264	4816.811
6	-2387.302	4788.603	4817.243
7	-2386.66	4789.32	4822.051
8	-2386.12	4790.241	4827.062
9	-2386.007	4792.015	4832.928
10	-2385.993	4793.986	4838.99

```
predictors for each model:
```

```
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 6-predictor model on the 442-patient dataset, we still find some high variance-inflation factors between the first and second serum variables. Note that they are far lower in magnitude than 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 5-predictor model, which 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	

2.2 Poisson model for doctor visits

The doctor visits dataset was examined in Cameron and Trivedi (1986). It contains data on the number of doctor visits in the past 2 weeks for a sample of 5,190 adults in the Australian Health Survey 1977–1978 (Australian Bureau of Statistics 1978). We will model the number of doctor visits using a Poisson regression on 12 predictors, which include gender, mean-adjusted age, age squared, income, and insurance status. Private insurance is indicated by the binary variable `private`. Insurance provided by the government because of low income is indicated by the binary variable `inslow`. Insurance provided by the government for retirees, the disabled, and veterans is indicated by the binary variable `insrdv`. The variable `hscore` is a health questionnaire score. The number of illnesses reported in the last week (`illness`) and whether the individual has had days of reduced activity because of illness or injury (`daysred`) are also included as predictors. Finally, indicators of whether an individual has a chronic condition that either limits his or her activity (`chcondlim`) or does not (`chcondnlim`) are included as predictors.

We estimate the parameters of the full model with `poisson` and then use the `mmp` command to draw marginal model plots to check the model's validity. The mean prediction for `poisson`, `n` is specified in `predict()`. We use a lowess estimate of the conditional mean by specifying `lowess` in the `smoother()` option. We draw a marginal model plot for the linear form and all continuous predictors by specifying `linear` and `predictors`, respectively.

```
. use docvisits
(Doctor visits)
. poisson docvis sex age agesq income private inslow insrdv illness hscore
> chcondlim chcondnlim daysred
Iteration 0: log likelihood = -3358.2094
Iteration 1: log likelihood = -3310.9148
Iteration 2: log likelihood = -3310.8564
Iteration 3: log likelihood = -3310.8564
```


(observed number of doctor visits) on the vertical axis variable. The closeness of the model and alternative fit lines corroborates the validity of the full model.

Now, we will use `gvselect` to fit the optimal model.

```
. gvselect <term> sex age agesq income private inslow insrdv illness hscore
> chcondlim chcondnlm daysred: poisson docvis <term>
```

Optimal models:

# Preds	LL	AIC	BIC
1	-3482.672	6969.344	6982.453
2	-3394.773	6795.546	6815.21
3	-3343.585	6695.169	6721.387
4	-3326.782	6663.564	6696.336
5	-3322.142	6656.285	6695.611
6	-3317.884	6649.768	6695.649
7	-3315.092	6646.185	6698.621
8	-3313.097	6644.194	6703.185
9	-3312.132	6644.263	6709.808
10	-3311.671	6645.343	6717.442
11	-3310.857	6645.713	6724.367
12	-3310.856	6647.713	6732.921

predictors for each model:

```
1 : daysred
2 : daysred illness
3 : daysred illness age
4 : daysred illness hscore age
5 : daysred illness hscore sex age
6 : daysred illness hscore inslow sex age
7 : daysred illness hscore inslow income sex age
8 : daysred illness hscore inslow chcondlim income sex age
9 : daysred illness hscore inslow chcondlim income sex age chcondnlm
10 : daysred illness hscore inslow chcondlim income sex age insrdv private
11 : daysred illness hscore inslow chcondlim income sex age insrdv private
    chcondnlm
12 : daysred illness hscore inslow chcondlim income sex age insrdv private
    chcondnlm agesq
```

The model with 8 predictors is favored by AIC, while the model with 5 predictors is favored by BIC. The AIC model includes the variables of the BIC model, income, the indicator for chronic conditions with activity limitation, `chcondlim`, and the indicator for government health insurance for those with low income, `inslow`.

We will examine the effect of the 8-predictor model chosen by AIC on the dataset with 5,190 adults. Recalling our discussion of cross-validation, we do not show the results of the model fit in `poisson`, because the significance levels would be misleading.

```
. quietly poisson docvis daysred illness hscore inslow chcondlim income sex age
. mmp, mean(n) smoother(lowess) linear predictors
```

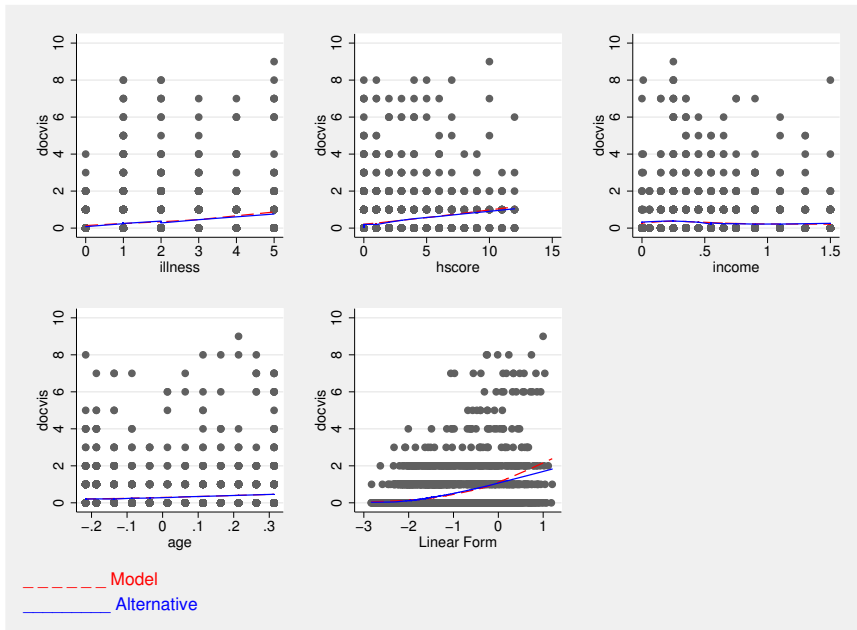


Figure 3. Marginal model plots for 8-predictor model

The marginal model plots in figure 3 support the validity of the 8-predictor model.

Now, we examine the 5-predictor model chosen by BIC.

```
. quietly poisson docvis daysred illness hscore sex age
. mmp, mean(n) smoother(lowess) linear predictors
```

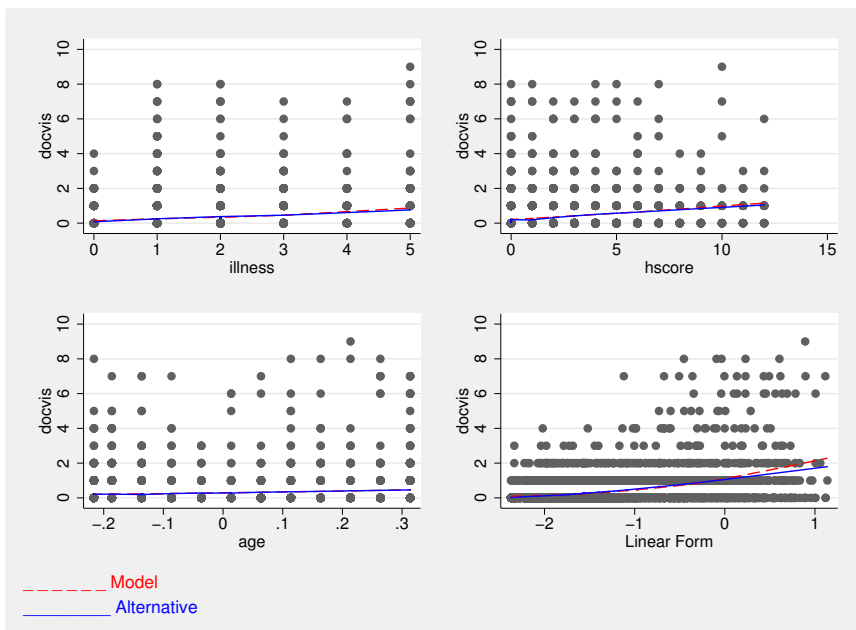


Figure 4. Marginal model plots for 5-predictor model

The marginal model plots in figure 4 support the validity of the model. If one does not favor BIC over AIC, or vice versa, you could choose this model for its parsimony.

3 Conclusion

We explored both the theory and practice of best subsets variable selection in regression. Using real datasets, we have demonstrated the use of the leaps-and-bounds algorithm in selecting regressors in normal and nonnormal models.

We fully defined the `gvselect` command as a method for performing regression variable selection in Stata and demonstrated its use with two different datasets.

4 References

- Agresti, A. 2013. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: Wiley.
- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.

- Australian Bureau of Statistics. 1978. Australian Health Survey 1977–78. Canberra, Australia: Australian Bureau of Statistics.
- Cameron, A. C., and P. K. Trivedi. 1986. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* 1: 29–53.
- Cook, R. D., and S. Weisberg. 1997. Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association* 92: 490–499.
- Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression. *Annals of Statistics* 32: 407–499.
- Furnival, G. M., and R. W. Wilson. 1974. Regressions by leaps and bounds. *Technometrics* 16: 499–511.
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.
- Hosmer, D. W., Jr., S. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.
- Hurvich, C. M., and C.-L. 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.
- Lindsey, C., and S. Sheather. 2010a. Model fit assessment via marginal model plots. *Stata Journal* 10: 215–225.
- . 2010b. Variable selection in linear regression. *Stata Journal* 10: 650–669.
- Ni, X., and X. Huo. 2005. Enhanced leaps-and-bounds method in subset selections with additional optimality tests.
https://www.informs.org/content/download/55245/522655/file/enhanced_leaps_and_bounds.pdf.
- 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.

About the authors

Charles Lindsey has a PhD from the Department of Statistics at Texas A&M University and is a senior statistician and software developer at StataCorp.

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