Description

varsoc reports the final prediction error (FPE), Akaike's information criterion (AIC), Schwarz's Bayesian information criterion (BIC), and the Hannan and Quinn information criterion (HQIC) lag-order selection statistics for a series of vector autoregressions of order 1 through a requested maximum lag. A sequence of likelihood-ratio test statistics for all the full vector autoregressive (VAR) models of order less than or equal to the highest lag order is also reported.

varsoc can be used as a preestimation or a postestimation command. The preestimation version can be used to select the lag order for a VAR model or vector error-correction (VEC) model. The postestimation version obtains the information needed to compute the statistics from the previous model or specified stored estimates.

Quick start

Compute AIC, BIC, and HQIC, and final prediction error to aid in the lag-order selection before VAR or VEC estimation of y1 and y2 using tsset data

    varsoc y1 y2

Same as above, but set the maximum lag order to be tested to 7

    varsoc y1 y2, maxlag(7)

Same as above, but use Lütkepohl's version of the information criteria

    varsoc y1 y2, maxlag(7) lutstats

Menu

Preestimation for VAR models

Statistics > Multivariate time series > VAR diagnostics and tests > Lag-order selection statistics (preestimation)

Postestimation for VAR models

Statistics > Multivariate time series > VAR diagnostics and tests > Lag-order selection statistics (postestimation)

Preestimation for VEC models

Statistics > Multivariate time series > VEC diagnostics and tests > Lag-order selection statistics (preestimation)

Postestimation for VEC models

Statistics > Multivariate time series > VEC diagnostics and tests > Lag-order selection statistics (postestimation)

Syntax

Preestimation syntax

varsoc *depvarlist* [*if*] [*in*] [, *preestimation_options*]

Postestimation syntax

varsoc [, <u>est</u>imates(*estname*)]

preestimation_options	Description
Main	
<u>max</u>lag(*#*)	set maximum lag order to *#*; default is maxlag(4)
<u>exog</u>(*varlist*)	use *varlist* as exogenous variables
<u>c</u>onstraints(*constraints*)	apply constraints to exogenous variables
noconstant	suppress constant term
lutstats	use Lütkepohl's version of information criteria
level(*#*)	set confidence level; default is level(95)
<u>sep</u>arator(*#*)	draw separator line after every *#* rows

You must tsset your data before using varsoc; see [TS] **tsset**.

by and collect are allowed with the preestimation version of varsoc; see [U] **11.1.10 Prefix commands**.

Preestimation options

Main

<u>max</u>lag(*#*) specifies the maximum lag order for which the statistics are to be obtained.

exog(*varlist*) specifies exogenous variables to include in the VAR models fit by varsoc.

constraints(*constraints*) specifies a list of constraints on the exogenous variables to be applied. Do not specify constraints on the lags of the endogenous variables because specifying one would mean that at least one of the VAR models considered by varsoc will not contain the lag specified in the constraint. Use var directly to obtain selection-order criteria with constraints on lags of the endogenous variables.

noconstant suppresses the constant terms from the model. By default, constant terms are included.

lutstats specifies that the Lütkepohl (2005) versions of the information criteria be reported. See *Methods and formulas* for a discussion of these statistics.

level(*#*) specifies the confidence level, as a percentage, that is used to identify the first likelihood-ratio test that rejects the null hypothesis that the additional parameters from adding a lag are jointly zero. The default is level(95) or as set by set level; see [U] **20.8 Specifying the width of confidence intervals**.

<u>sep</u>arator(*#*) specifies how often separator lines should be drawn between rows. By default, separator lines do not appear. For example, separator(1) would draw a line between each row, separator(2) between every other row, and so on.

Postestimation option

estimates(*estname*) specifies the name of a previously stored set of var, svar, or ivsvar estimates. When no *depvarlist* is specified, varsoc uses the *postestimation syntax* and uses the currently active estimation results or the results specified in estimates(*estname*). See [R] **estimates** for information on manipulating estimation results.

Remarks and examples

Many selection-order statistics have been developed to assist researchers in fitting a VAR model of the correct order. Several of these selection-order statistics appear in the [TS] **var** output. The varsoc command computes these statistics over a range of lags p while maintaining a common sample and option specification.

varsoc can be used as a preestimation or a postestimation command. When it is used as a preestimation command, a *depvarlist* is required, and the default maximum lag is 4. When it is used as a postestimation command, varsoc uses the model specification stored in *estname* or the previously fitted model.

varsoc computes four information criteria as well as a sequence of likelihood ratio (LR) tests. The information criteria include the FPE, AIC, the HQIC, and BIC.

For a given lag p, the LR test compares a VAR model with p lags with one with $p - 1$ lags. The null hypothesis is that all the coefficients on the pth lags of the endogenous variables are zero. To use this sequence of LR tests to select a lag order, we start by looking at the results of the test for the model with the most lags, which is at the bottom of the table. Proceeding up the table, the first test that rejects the null hypothesis is the lag order selected by this process. See Lütkepohl (2005, 143–144) for more information on this procedure. An '*' appears next to the LR statistic indicating the optimal lag.

For the remaining statistics, the lag with the smallest value is the order selected by that criterion. An '*' indicates the optimal lag. Strictly speaking, the FPE is not an information criterion, though we include it in this discussion because, as with an information criterion, we select the lag length corresponding to the lowest value; and, naturally, we want to minimize the prediction error. The AIC measures the discrepancy between the given model and the true model, which, of course, we want to minimize. Amemiya (1985) provides an intuitive discussion of the arguments in Akaike (1973). The BIC and the HQIC can be interpreted similarly to the AIC, though the BIC and the HQIC have a theoretical advantage over the AIC and the FPE. As Lütkepohl (2005, 148–152) demonstrates, choosing p to minimize the BIC or the HQIC provides consistent estimates of the true lag order, p. In contrast, minimizing the AIC or the FPE will overestimate the true lag order with positive probability, even with an infinite sample size.

Although VAR models assume that the modulus is strictly less than 1 (see [TS] **varstable**), VEC models do not need to satisfy this condition, and they work even if all the variables included in the model are integrated of order 1, I(1). Regardless of these differences, varsoc works for both estimation commands. As shown by Nielsen (2006), the lag-order selection statistics discussed above can be used in the presence of I(1) variables.

▷ Example 1: Preestimation

Here we use `varsoc` as a preestimation command.

```
. use https://www.stata-press.com/data/r19/lutkepohl2
(Quarterly SA West German macro data, Bil DM, from Lutkepohl 1993 Table E.1)

. varsoc dln_inv dln_inc dln_consump if qtr<=tq(1978q4), lutstats

Lutkepohl's lag-order selection criteria
     Sample: 1961q2 thru 1978q4                       Number of obs = 71

  ┌──────────────────────────────────────────────────────────────────────────┐
  │ Lag │    LL       LR      df    p      FPE       AIC       HQIC      SBIC   │
  ├──────────────────────────────────────────────────────────────────────────┤
  │  0  │ 564.784                         2.7e-11  -24.423   -24.423*  -24.423* │
  │  1  │ 576.409  23.249    9  0.006 2.5e-11   -24.497  -24.3829  -24.2102     │
  │  2  │ 588.859  24.901*   9  0.003 2.3e-11* -24.5942* -24.3661  -24.0205     │
  │  3  │ 591.237   4.7566   9  0.855 2.7e-11  -24.4076  -24.0655  -23.5472     │
  │  4  │ 598.457  14.438    9  0.108 2.9e-11  -24.3575  -23.9012  -23.2102     │
  └──────────────────────────────────────────────────────────────────────────┘

  * optimal lag
  Endogenous: dln_inv dln_inc dln_consump
   Exogenous: _cons
```

The sample used begins in 1961q2 because all the VAR models are fit to the sample defined by any `if` or `in` conditions and the available data for the maximum lag specified. The default maximum number of lags is four. Because we specified the `lutstats` option, the table contains the Lütkepohl (2005) versions of the information criteria, which differ from the standard definitions in that they drop the constant term from the log likelihood. In this example, the likelihood-ratio tests selected a model with two lags. AIC and FPE have also both chosen a model with two lags, whereas BIC and HQIC have both selected a model with zero lags.

◁

▷ Example 2: Postestimation

`varsoc` works as a postestimation command when no dependent variables are specified.

```
. var dln_inc dln_consump if qtr<=tq(1978q4), lutstats exog(l.dln_inv)
  (output omitted)

. varsoc

Lutkepohl's lag-order selection criteria
     Sample: 1960q4 thru 1978q4                       Number of obs = 73

  ┌──────────────────────────────────────────────────────────────────────────┐
  │ Lag │    LL       LR      df    p      FPE       AIC       HQIC      SBIC   │
  ├──────────────────────────────────────────────────────────────────────────┤
  │  0  │ 460.646                         1.3e-08  -18.2962  -18.2962  -18.2962*│
  │  1  │ 467.606  13.919    4  0.008 1.2e-08   -18.3773  -18.3273  -18.2518   │
  │  2  │ 477.087  18.962*   4  0.001 1.0e-08* -18.5275* -18.4274* -18.2764    │
  └──────────────────────────────────────────────────────────────────────────┘

  * optimal lag
  Endogenous: dln_inc dln_consump
   Exogenous: L.dln_inv  _cons
```

Because we included one lag of `dln_inv` in our original model, `varsoc` did likewise with each model it fit.

◁

Based on the work of Tsay (1984), Paulsen (1984), and Nielsen (2006), these lag-order selection criteria can be used to determine the lag length of the VAR model underlying a VEC model. See [TS] **vec intro** for an example in which we use varsoc to choose the lag order for a VEC model.

Stored results

varsoc stores the following in r():

Scalars
r(N)	number of observations
r(tmax)	last time period in sample
r(tmin)	first time period in sample
r(mlag)	maximum lag order
r(N_gaps)	the number of gaps in the sample

Macros
r(endog)	names of endogenous variables
r(lutstats)	lutstats, if specified
r(cns#)	the #th constraint
r(exog)	names of exogenous variables
r(rmlutstats)	rmlutstats, if specified

Matrices
r(stats)	LL, LR, FPE, AIC, HQIC, BIC, and p-values

Methods and formulas

Methods and formulas are presented under the following headings:

> Likelihood-ratio statistic
> Model-order statistics
> Lutstats

Likelihood-ratio statistic

As shown by Hamilton (1994, 295–296), the log likelihood for a VAR(p) model is

$$\text{LL} = \left(\frac{T}{2}\right)\left\{\ln(|\widehat{\boldsymbol{\Sigma}}^{-1}|) - K\ln(2\pi) - K\right\}$$

where T is the number of observations, K is the number of equations, and $\widehat{\boldsymbol{\Sigma}}$ is the maximum likelihood estimate of $E[\mathbf{u}_t\mathbf{u}_t']$, where $\mathbf{u}_t$ is the $K \times 1$ vector of disturbances. Because

$$\ln(|\widehat{\boldsymbol{\Sigma}}^{-1}|) = -\ln(|\widehat{\boldsymbol{\Sigma}}|)$$

the log likelihood can be rewritten as

$$\text{LL} = -\left(\frac{T}{2}\right)\left\{\ln(|\widehat{\boldsymbol{\Sigma}}|) + K\ln(2\pi) + K\right\}$$

Letting LL(j) be the value of the log likelihood with j lags yields the LR statistic for lag order j as

$$\text{LR}(j) = 2\{\text{LL}(j) - \text{LL}(j-1)\}$$

Model-order statistics

The formula for the FPE given in Lütkepohl (2005, 147) is

$$\text{FPE} = |\mathbf{\Sigma}_u| \left(\frac{T + Kp + 1}{T - Kp - 1} \right)^K$$

This formula, however, assumes that there is a constant in the model and that none of the variables are omitted because of collinearity. To deal with these problems, the FPE is implemented as

$$\text{FPE} = |\mathbf{\Sigma}_u| \left(\frac{T + \overline{m}}{T - \overline{m}} \right)^K$$

where $\overline{m}$ is the average number of parameters over the K equations. This implementation accounts for variables omitted because of collinearity.

By default, the AIC, BIC, and HQIC are computed according to their standard definitions, which include the constant term from the log likelihood. That is,

$$\text{AIC} = -2 \left(\frac{\text{LL}}{T} \right) + \frac{2t_p}{T}$$

$$\text{BIC} = -2 \left(\frac{\text{LL}}{T} \right) + \frac{\ln(T)}{T} t_p$$

$$\text{HQIC} = -2 \left(\frac{\text{LL}}{T} \right) + \frac{2\ln\{\ln(T)\}}{T} t_p$$

where t_p is the total number of parameters in the model and LL is the log likelihood.

Lutstats

Lütkepohl (2005) advocates dropping the constant term from the log likelihood because it does not affect inference. The Lütkepohl versions of the information criteria are

$$\text{AIC} = \ln(|\mathbf{\Sigma}_u|) + \frac{2pK^2}{T}$$

$$\text{BIC} = \ln(|\mathbf{\Sigma}_u|) + \frac{\ln(T)}{T} pK^2$$

$$\text{HQIC} = \ln(|\mathbf{\Sigma}_u|) + \frac{2\ln\{\ln(T)\}}{T} pK^2$$

References

Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle". In *Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki, 267–281. Budapest: Akailseoniai–Kiudo.

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Hamilton, J. D. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press. https://doi.org/10.2307/j. ctv14jx6sm.

Lütkepohl, H. 1993. *Introduction to Multiple Time Series Analysis*. 2nd ed. New York: Springer.

———. 2005. *New Introduction to Multiple Time Series Analysis*. New York: Springer.

Nielsen, B. 2006. "Order determination in general vector autoregressions". In *Time Series and Related Topics*. Lecture Notes—Monograph Series, edited by H.-C. Ho, C.-K. Ing, and T. L. Lai, vol. 52: 93–112. Beachwood, OH: Institute of Mathematical Statistics. https://doi.org/10.1214/074921706000000978.

Paulsen, J. 1984. Order determination of multivariate autoregressive time series with unit roots. *Journal of Time Series Analysis* 5: 115–127. https://doi.org/10.1111/j.1467-9892.1984.tb00381.x.

Schenck, D. 2016. Vector autoregressions in Stata. *The Stata Blog: Not Elsewhere Classified.* https://blog.stata.com/2016/08/09/vector-autoregressions-in-stata/.

Tsay, R. S. 1984. Order selection in nonstationary autoregressive models. *Annals of Statistics* 12: 1425–1433. https://doi.org/10.1214/aos/1176346801.

Also see

[TS] **var** — Vector autoregressive models

[TS] **var intro** — Introduction to vector autoregressive models

[TS] **var ivsvar** — Instrumental-variables structural vector autoregressive models

[TS] **var svar** — Structural vector autoregressive models

[TS] **varbasic** — Fit a simple VAR and graph IRFs or FEVDs

[TS] **vec** — Vector error-correction models

[TS] **vec intro** — Introduction to vector error-correction models