**arimasoc** — Obtain lag-order selection statistics for ARMAs



# Description

arimasoc reports Akaike's information criterion (AIC), Schwarz's Bayesian information criterion (BIC), and the Hannan and Quinn information criterion (HQIC) for a series of autoregressive moving-average (ARMA) models. These criteria are used to select the number of autoregressive (AR) and moving-average (MA) lags to be used in the ARMA model.

# Quick start

Compute AIC, BIC, and HQIC for ARMA models of y with up to 2 AR and 2 MA lags

```
arimasoc y
```

Same as above, but compare ARMA models of y with up to 7 AR lags and 1 MA lag

```
arimasoc y, maxar(7) maxma(1)
```

Pass the option condition to the underlying arima estimation

```
arimasoc y, arimaopts(condition)
```

Limit the number of iterations in arima estimation to 50

```
arimasoc y, arimaopts(iterate(50))
```

Compute AIC, BIC, and HQIC for ARMAX models of y with exogenous regressor x, and with up to 2 AR lags and 2 MA lags

```
arimasoc y x
```

Compute information criteria for ARMA models in the difference of y

```
arimasoc d.y
```

# Menu

Statistics > Time series > ARIMA and ARMAX > Lag-order selection

## Syntax

`arimasoc` *depvar* [ *indepvars* ] [ *if* ] [ *in* ] [ , *options* ]

| *options* | Description |
|---|---|
| Main | |
| `maxar(#)` | set maximum AR order to #; default is `maxar(2)` |
| `maxma(#)` | set maximum MA order to #; default is `maxma(2)` |
| `n(#)` | use $N$ when calculating BIC and HQIC |
| `arimaopts(`*opts*`)` | specify options of `arima` for model estimation |

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

*depvar* and *indepvars* may contain time-series operators; see [U] **11.4.4 Time-series varlists**.

## Options

Main

`maxar(#)` specifies the maximum AR lag order for which the information criteria are to be calculated. The default is `maxar(2)`.

`maxma(#)` specifies the maximum MA lag order for which the information criteria are to be calculated. The default is `maxma(2)`.

`n(#)` sets $N$ to be used when calculating BIC and HQIC; see [R] **IC note**.

`arimaopts(`*opts*`)` specifies options of `arima` to include in the ARMAs fit by `arimasoc`. *opts* may be `noconstant`, `sarima()`, `mar()`, `mma()`, `condition`, `savespace`, `diffuse`, `difficult`, `technique()`, `iterate()`, `tolerance()`, `ltolerance()`, `nrtolerance()`, `gtolerance()`, `nonrtolerance`, and `collinear`. See [TS] **arima** for a description of these options.

## Remarks and examples

Information criteria are statistics that help researchers choose the best model. In general, these statistics balance goodness of fit against parsimony. The statistics decrease with the model's goodness of fit, as assessed by the likelihood function, and increase with the number of parameters. Therefore, the selected model is the one that minimizes the information criterion, or equivalently, the model that best fits the data while using the fewest parameters possible.

`arimasoc` computes three information criteria (AIC, BIC, and HQIC) that help researchers determine the correct number of AR and MA lags to be included in an ARMA model. `arimasoc` calculates these criteria for ARMA models with up to $p$ AR lags and $q$ MA lags, where $p$ and $q$ are predetermined numbers. `arimasoc` keeps the sample and option specifications the same in the estimation of all the different ARMA models.

Different information criteria may choose different models. Among the three information criteria available, the BIC and HQIC have the advantage that they are consistent. This means that as the sample size grows, they select the correct number of lags with probability approaching one. However, there is a positive probability that AIC will select more lags than necessary, even with an infinite sample size. See Brockwell and Davis (2016, 149–151).

### ▷ Example 1: Basic example

We use `arimasoc` on US macro data to fit several ARMA models of the output gap.

```
. use https://www.stata-press.com/data/r19/usmacro
(Federal Reserve Economic Data - St. Louis Fed)

. arimasoc ogap
Fitting models (9): ......... done

Lag-order selection criteria

Sample: 1954q3 thru 2010q4                          Number of obs = 226
```

| Model | LL | df | AIC | BIC | HQIC |
|---|---|---|---|---|---|
| ARMA(0,0) | -549.4036 | 2 | 1102.807 | 1109.648 | 1105.568 |
| ARMA(0,1) | -435.0753 | 3 | 876.1507 | 886.4123 | 880.2919 |
| ARMA(0,2) | -361.249 | 4 | 730.4981 | 744.1802 | 736.0196 |
| ARMA(1,0) | -292.3313 | 3 | 590.6625 | 600.9241 | 594.8037 |
| ARMA(1,1) | -281.5762 | 4 | 571.1524 | 584.8345 | 576.6739 |
| ARMA(1,2) | -275.3697 | 5 | 560.7395 | 577.8422 | 567.6414 |
| ARMA(2,0) | -276.5956 | 4 | 561.1912 | 574.8733 | 566.7127 |
| ARMA(2,1) | -273.9052 | 5 | 557.8104 | 574.9131 | 564.7123 |
| ARMA(2,2) | -273.2799 | 6 | 558.5599 | 579.0831 | 566.8422 |

```
Selected (max) LL:   ARMA(2,2)
Selected (min) AIC:  ARMA(2,1)
Selected (min) BIC:  ARMA(2,0)
Selected (min) HQIC: ARMA(2,1)
```

The default maximum AR lag $p$ and MA lag $q$ are both 2. The table provides results for each AR and MA combination, beginning with a constant-only model `ARMA(0,0)`. The column `LL` reports the log likelihood, and the column `df` reports the number of estimated parameters. In this example, the log likelihood is maximized with the most complex model, `ARMA(2,2)`. AIC and HQIC select a model with two AR lags and one MA lag. BIC has selected a model with two AR lags and no MA terms. Model selection criteria can disagree because they put different penalties on the complexity of the model, as measured by the number of parameters estimated. All selection criteria here suggest two AR lags, but they differ in their recommendation of MA terms.

◁

### ▷ Example 2: Changing the maximum order

Continuing with example 1, we want to change the maximum lag order for both AR and MA. We will use the `maxar()` and `maxma()` options for this. We include up to three AR lags and up to one MA lag.

```
. arimasoc ogap, maxar(3) maxma(1)
Fitting models (8): ........ done
Lag-order selection criteria
Sample: 1954q3 thru 2010q4                         Number of obs = 226
```

| Model | LL | df | AIC | BIC | HQIC |
|---|---|---|---|---|---|
| ARMA(0,0) | -549.4036 | 2 | 1102.807 | 1109.648 | 1105.568 |
| ARMA(0,1) | -435.0753 | 3 | 876.1507 | 886.4123 | 880.2919 |
| ARMA(1,0) | -292.3313 | 3 | 590.6625 | 600.9241 | 594.8037 |
| ARMA(1,1) | -281.5762 | 4 | 571.1524 | 584.8345 | 576.6739 |
| ARMA(2,0) | -276.5956 | 4 | 561.1912 | 574.8733 | 566.7127 |
| ARMA(2,1) | -273.9052 | 5 | 557.8104 | 574.9131 | 564.7123 |
| ARMA(3,0) | -273.2421 | 5 | 556.4843 | 573.587 | 563.3862 |
| ARMA(3,1) | -273.1883 | 6 | 558.3766 | 578.8998 | 566.6589 |

```
Selected (max) LL:   ARMA(3,1)
Selected (min) AIC:  ARMA(3,0)
Selected (min) BIC:  ARMA(3,0)
Selected (min) HQIC: ARMA(3,0)
```

Again, the log likelihood is maximized with the most complex model, using the full three AR lags and one MA lag. The information criteria all select the model with three AR lags and zero MA lags. Because the selected model is on the boundary of the range of AR terms, it may be worthwhile to consider an even larger range of AR terms to ensure that we have found the true minimum of the information criteria.

◁

## Stored results

`arimasoc` stores the following in `r()`:

Scalars

| | |
|---|---|
| `r(N)` | number of observations |
| `r(ar_max)` | maximum AR lag order |
| `r(ma_max)` | maximum MA lag order |

Macros

| | |
|---|---|
| `r(depvar)` | name of endogenous variable |
| `r(covariates)` | names of exogenous variables |
| `r(aic_sel)` | selected ARMA model by AIC |
| `r(bic_sel)` | selected ARMA model by BIC |
| `r(hqic_sel)` | selected ARMA model by HQIC |
| `r(ll_sel)` | selected ARMA model by LL |
| `r(aic_cmd)` | selected ARMA command by AIC |
| `r(bic_cmd)` | selected ARMA command by BIC |
| `r(hqic_cmd)` | selected ARMA command by HQIC |
| `r(ll_cmd)` | selected ARMA command by LL |

Matrices

| | |
|---|---|
| `r(table)` | table of results |
| `r(converged)` | 1 if converged, 0 otherwise |

## Methods and formulas

Akaike's (1974) information criterion is defined as

$$\text{AIC} = -2\ln L + 2k$$

where $\ln L$ is the maximized log likelihood of the model and $k$ is the number of parameters estimated. Some authors define AIC as the expression above divided by the sample size.

Schwarz's (1978) Bayesian information criterion is another measure of fit. It is defined as

$$\text{BIC} = -2\ln L + k\ln N$$

where $N$ is the sample size. See [R] **IC note** for additional information on calculating and interpreting BIC.

The Hannan and Quinn (1979) information criterion is another measure of fit. It is defined as

$$\text{HQIC} = -2\ln L + 2k\ln\ln N$$

## References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Brockwell, P. J., and R. A. Davis. 2016. *Introduction to Time Series and Forecasting*. 3rd ed. Cham, Switzerland: Springer.

Hannan, E. J., and B. G. Quinn. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, B ser., 41: 190–195. https://doi.org/10.1111/j.2517-6161.1979.tb01072.x.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464. https://doi.org/10.1214/aos/1176344136.

## Also see

[TS] **arima** — ARIMA, ARMAX, and other dynamic regression models

[TS] **arfimasoc** — Obtain lag-order selection statistics for ARFIMAs

[TS] **varsoc** — Obtain lag-order selection statistics for VAR and VEC models



For suggested citations, see the FAQ on citing Stata documentation.