Title

stata.com

IC note — Calculating and interpreting information criteria

Description

This entry discusses a statistical issue that arises when using the Bayesian (BIC), consistent Akaike's (CAIC), and corrected Akaike's (AICc) information criteria to compare models.

Stata calculates BIC, CAIC, and AICc using $N =$ `e(N)`, unless `e(N_ic)` has been set; in that instance, it uses $N =$ `e(N_ic)`. For example, choice-model `cm` commands set `e(N_ic)` to the number of cases because these commands use a data arrangement in which multiple Stata observations represent a single statistical observation, which is called a case.

Sometimes, it would be better if a different N than `e(N)` were used. Commands that calculate BIC, CAIC, and AICc have an `n()` option, allowing you to specify the N to be used.

In summary,

1. if you are comparing results estimated by the same estimation command, using the default BIC, CAIC, or AICc calculation is probably fine. There is an issue, but most researchers would ignore it.
2. if you are comparing results estimated by different estimation commands, you need to be on your guard.
    a. If the different estimation commands share the same definitions of observations, independence, and the like, you are back to case 1.
    b. If they differ in these regards, you need to think about the value of N that should be used. For example, `logit` and `xtlogit` differ in that the former assumes independent observations and the latter, independent panels.
    c. If estimation commands differ in the events being used over which the likelihood function is calculated, the information criteria may not be comparable at all. We say information *criteria* because this would apply equally to the Akaike information criterion (AIC) and its possible extensions AICc and CAIC, as well as to the BIC. For instance, `streg` and `stcox` produce such incomparable results. The events used by `streg` are the actual survival times, whereas the events used by `stcox` are failures within risk pools, conditional on the times at which failures occurred.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Background

The AIC and the BIC are two popular measures for comparing maximum likelihood models. AIC and BIC are defined as

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

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

where

$$\ln L = \text{maximized log-likelihood}$$

$$k = \text{number of parameters estimated}$$

$$N = \text{number of observations}$$

However, when sample size is small, AIC is biased, and Burnham and Anderson (2002) suggest to use AICc,

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{N-k-1}$$

CAIC is a consistent version of AIC and was proposed in Bozdogan (1987),

$$\text{CAIC} = -2\ln L + k(\ln N + 1)$$

All four information criteria—AIC, BIC, CAIC, and AICc—can be viewed as measures that combine fit and complexity. Fit is measured negatively by $-2\ln L$; the larger the value, the worse the fit. Complexity is measured positively, for example, by $2k$ (AIC) or $k\ln N$ (BIC).

Given two models fit on the same data, the model with the smaller value of the information criterion is considered to be better.

There is substantial literature on these measures: see Akaike (1974); Raftery (1995); Sakamoto, Ishiguro, and Kitagawa (1986); Schwarz (1978); Burnham and Anderson (2002); and Hurvich and Tsai (1989).

When Stata calculates the above measures, it uses the rank of `e(V)` for k, and it uses `e(N)` for N. `e(V)` and `e(N)` are Stata notation for results stored by the estimation command. `e(V)` is the variance–covariance matrix of the estimated parameters, and `e(N)` is the number of observations in the dataset used in calculating the result.

The problem of determining N

The difference between AIC and the other three information criteria is that AIC uses the constant 2 to weight k, whereas the complexity term for BIC, CAIC, and AICc depends on N.

Determining what value of N should be used is problematic. Despite appearances, the definition "N is the number of observations" is not easy to make operational. N does not appear in the likelihood function itself, N is not the output of a standard statistical formula, and what is an observation is often subjective.

Example 1

Often what is meant by N is obvious. Consider a simple logit model. What is meant by N is the number of observations that is statistically independent and that corresponds to M, the number of observations in the dataset used in the calculation. We will write $N = M$.

But now assume that the same dataset has a grouping variable and the data are thought to be clustered within group. To keep the problem simple, let's pretend that there are G groups and m observations within group, so that $M = G \times m$. Because you are worried about intragroup correlation, you fit your model with `xtlogit`, grouping on the grouping variable. Now, you wish to calculate BIC. What is the N that should be used? $N = M$ or $N = G$?

That is a deep question. If the observations really are independent, then you should use $N = M$. If the observations within group are not just correlated but are duplicates of one another, and they had to be so, then you should use $N = G$ (Kass and Raftery 1995). Between those two extremes, you should probably use a number between N and G, but determining what that number should be from measured correlations is difficult. Using $N = M$ is conservative in that, if anything, it overweights complexity. Conservativeness, however, is subjective, too: using $N = G$ could be considered more conservative in that fewer constraints are being placed on the data.

When the estimated correlation is high, our reaction would be that using $N = G$ is probably more reasonable. Our first reaction, however, would be that using BIC to compare models is probably a misuse of the measure.

Stata uses $N = M$. An informal survey of web-based literature suggests that $N = M$ is the popular choice.

There is another reason, not so good, to choose $N = M$. It makes across-model comparisons more likely to be valid when performed without thinking about the issue. Say that you wish to compare the `logit` and `xtlogit` results. Thus, you need to calculate

$$\text{BIC}_p = -2\ln L_p + k\ln N_p$$

$$\text{BIC}_x = -2\ln L_x + k\ln N_x$$

Whatever N you use, you must use the same N in both formulas. Stata's choice of $N = M$ at least meets that test.

◁

Example 2

In the above example, using $N = M$ is reasonable. Now, let's look at when using $N = M$ is wrong, even if popular.

Consider a model fit by `stcox`. Using $N = M$ is certainly wrong if for no other reason than M is not even a well-defined number. The same data can be represented by different datasets with different numbers of observations. For example, in one dataset, there might be one observation per subject. In another, the same subjects could have two records each, the first recording the first half of the time at risk and the second recording the remaining part. All statistics calculated by Stata on either dataset would be the same, but M would be different.

Deciding on the right definition, however, is difficult. Viewed one way, N in the Cox regression case should be the number of risk pools, R, because the Cox regression calculation is made on the basis of the independent risk pools. Viewed another way, N should be the number of subjects, N_{subj}, because, even though the likelihood function is based on risk pools, the parameters estimated are at the subject level.

You can decide which argument you prefer.

For parametric survival models, in single-record data, $N = M$ is unambiguously correct. For multirecord data, there is an argument for $N = M$ and for $N = N_{\text{subj}}$.

◁

The problem of conformable likelihoods

The problem of conformable likelihoods does not concern N. Researchers sometimes use information criteria such as BIC and AIC to make comparisons across models. For that to be valid, the likelihoods must be conformable; that is, the likelihoods must all measure the same thing.

It is common to think of the likelihood function as the $\Pr(\text{data} \mid \text{parameters})$, but in fact, the likelihood is

$$\Pr(\text{particular events in the data} \mid \text{parameters})$$

You must ensure that the events are the same.

For instance, they are not the same in the semiparametric Cox regression and the various parametric survival models. In Cox regression, the events are, at each failure time, that the subjects observed to fail in fact failed, given that failures occurred at those times. In the parametric models, the events are that each subject failed exactly when the subject was observed to fail.

The formula for AIC, AICc, CAIC, and BIC can be written as

$$\text{measure} = -2\ln L + \text{complexity}$$

When you are comparing models, if the likelihoods are measuring different events, even if the models obtain estimates of the same parameters, differences in the information measures are irrelevant.

The first problem does not arise with AIC; the second problem does

Regardless of model, the problem of defining N never arises with AIC because N is not used in the AIC calculation. AIC uses a constant 2 to weight complexity as measured by k, rather than $\ln N$.

However, for all four information criteria—AIC, AICc, CAIC, and BIC—the likelihood functions must be conformable; that is, they must be measuring the same event.

Calculating BIC, AICc, and CAIC correctly

When using BIC, AICc, or CAIC to compare results, and especially when using them to compare results from different models, you should think carefully about how N should be defined. Then, specify that number by using the `n()` option:

```
. estimates stats full sub, all n(74)
Information criteria
```

Model	N	ll(null)	ll(model)	df
full	74	-45.03321	-20.59083	4
sub	74	-45.03321	-27.17516	3

Model	AIC	BIC	AICc	CAIC
full	49.18167	58.39793	49.76138	62.39793
sub	60.35031	67.26251	60.69317	70.26251

```
Legend: AIC is Akaike's information criterion.
        BIC is Bayesian information criterion.
        AICc is corrected Akaike's information criterion.
        CAIC is consistent Akaike's information criterion.
```

Both `estimates stats` and `estat ic` allow the `n()` option; see [R] **estimates stats** and [R] **estat ic**.

Methods and formulas

AIC, BIC, CAIC, and AICc are defined as

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

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

$$\text{CAIC} = -2\ln L + k(\ln N + 1)$$

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{N-k-1}$$

where $\ln L$ is the maximized log-likelihood of the model; k is the model degrees of freedom calculated as the rank of variance–covariance matrix of the parameters `e(V)`, unless the `df()` option is specified; and N is the number of observations used in estimation or, more precisely, the number of independent terms in the likelihood. Operationally, N is defined as `e(N)`, unless the estimation command returns `e(N_ic)` or the `n()` option is specified with `estimates stats` or `estat ic`.

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.

Bozdogan, H. 1987. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* 52: 345–370. https://doi.org/10.1007/BF02294361.

Burnham, K. P., and D. R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer.

Hurvich, C. M., and C.-L. Tsai. 1989. Regression and time series model selection in small samples. *Biometrika* 76: 297–307. https://doi.org/10.1093/biomet/76.2.297.

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90: 773–795. https://doi.org/10.1080/01621459.1995.10476572.

Raftery, A. E. 1995. Bayesian model selection in social research. In Vol. 25 of *Sociological Methodology*, ed. P. V. Marsden, 111–163. Oxford: Blackwell.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. *Akaike Information Criterion Statistics*. Dordrecht, The Netherlands: Reidel.

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

Also see

[R] **estat ic** — Display information criteria

[R] **estimates stats** — Model-selection statistics