Information Criteria for Model Selection

2018-02-11

We are going to use the log-likelihood for model selection. To do this, we fit a set of candidate models we are choosing between using ML estimation and compute the log-likelihood value for each model. Models with higher log-likelihood values will be favored; they are more likely given the data.

Remember that the value of log-likelihood is the same for both the ML and OLS estimated models. Thus, it does not matter how we fit the model; we can use `lm()` to estimate the coefficients from a model since it is easier to use.

Preparation

In this set of notes, we will use the data in the *edschools-2010.csv* file. These data include institutional-level attributes for several graduate education schools/programs rated in 2010 by *U.S. News and World Report*. The attributes include:

- `school`: Institution name
- `rank`: U.S. News and World Report rank
- `peer_rate`: Peer assessment score (5.0 = highest).
- `gre`: Average GRE score for 2009 incoming students
- `sf_ratio`: 2009 student-to-faculty ratio
- `phd_fac_ratio`: 2009 doctoral student-to-faculty ratio
- `avg_res`: 2009 funded research per faculty member (in thousands of dollars)
- `doc_accept`: 2009 doctoral acceptance rate

```
# Load libraries
library(AICcmodavg)
library(broom)
library(dplyr)
library(readr)

# Import the data
ed = read_csv(file = "~/Dropbox/epsy-8252/data/ed-schools-2010.csv")
head(ed)
```

school	rank	peer_rate	gre	sf_ratio	phd_fac_ratio	avg_res	doc_accept
Vanderbilt University (Peabody) (TN)	1	4.5	683	2.7	0.5	35.4	7.1
Teachers College, Columbia University (NY)	2	4.4	617	6.1	1.6	41.4	22.0
Harvard University (MA)	3	4.5	671	5.8	0.8	18.1	8.9
Stanford University (CA)	3	4.7	688	3.8	0.7	15.2	8.0
University of Oregon	5	3.5	550	4.1	1.3	29.9	13.4
Johns Hopkins University (MD)	6	3.9	612	0.3	0.1	19.3	30.6

Scientific Hypotheses

We are interested in understanding how our peers in education rate other programs. Based on the substantive literature we have three hypotheses about how programs are rated:

- **H1:** The rating of a program is attributable to the quality of its students.
- **H2:** Peer ratings are attributable to the level of interaction faculty members have with students.

- **H3:** Peer ratings are attributable to the research prestige of the programs.

Translating Hypotheses to Models

We need to translate these hypotheses into statistical models that we can then fit to a set of data. The models are only proxies for these hypotheses. However, that being said, the validity of using the models as proxies is dependent on whether we have measured well, whether the translation makes substantive sense given the literature base, etc. Here is how we are measuring the different attributes:

- We will use average GRE score as a measure of student-body quality.
- We will use student-to faculty ratio and doctoral student-to-faculty ratio as measures of faculty interaction.
- We will use doctoral student acceptance rate and average faculty research funding as measures of research prestige.

Once this has been identified, we can write out the models associated with the scientific hypotheses. These models using regression notation are:

- **M1:** $\text{Peer Rating}_i = \beta_0 + \beta_1(\text{GRE}_i) + \epsilon_i$
- **M2:** $\text{Peer Rating}_i = \beta_0 + \beta_1(\text{student-to-faculty ratio}_i) + \beta_2(\text{doctoral student-to-faculty ratio}_i) + \epsilon_i$
- **M3:** $\text{Peer Rating}_i = \beta_0 + \beta_1(\text{doctoral acceptance rate}_i) + \beta_2(\text{average research funding}_i) + \epsilon_i$

These are referred to as the *candidate models*. Now we fit these three candidate models in R.

```
lm.1 = lm(peer_rate ~ 1 + gre, data = ed)
lm.2 = lm(peer_rate ~ 1 + sf_ratio + phd_fac_ratio, data = ed)
lm.3 = lm(peer_rate ~ 1 + doc_accept + avg_res, data = ed)
```

Compute Log-Likelihood

Now we can compute the log-likelihood values for each model.

```
logLik(lm.1)
```

```
## 'log Lik.' -23.6 (df=3)
```

```
logLik(lm.2)
```

```
## 'log Lik.' -20.9 (df=4)
```

```
logLik(lm.3)
```

```
## 'log Lik.' -24.4 (df=4)
```

The log-likelihood values are also available from the `glance()` function's output.

```
glance(lm.1)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.219	0.203	0.388	14	0	2	-23.6	53.2	59	7.54	50

These values suggest that the best fitting model is Model 2; it has the highest log-likelihood value.

Deviance: An Alternative Fit Value

It is common to multiply the log-likelihood values by -2. This is called the *deviance*. This alleviates the deviance. Deviance is a measure of error, so when evaluating deviance values, lower is better. (The square brackets in the syntax grab the log-likelihood value from the `logLik()` output.)

```
-2 * logLik(lm.1)[1] #Model 1
```

```
## [1] 47.2
```

```
-2 * logLik(lm.2)[1] #Model 2
```

```
## [1] 41.9
```

```
-2 * logLik(lm.3)[1] #Model 3
```

```
## [1] 48.8
```

Again, the best fitting model is Model 2; it has the lowest deviance value. Whether you evaluate using the log-likelihood, or the deviance, you will end up with the same model. Using deviance, however, has the advantages of (1) having a direct relationship to model error, so it is more interpretable, and (2) not being negative.

Compute AIC Values

Remember that lower values of deviance indicate the model (as defined via the set of parameters) is more likely given the data. However, in practice we cannot directly compare the deviances since the models include a different number of parameters. To account for this, we will add a penalty term to the deviance,

$$AIC = \text{Deviance} + 2(k)$$

where k is the number of parameters being estimated in the model (including the intercept and RMSE). Note that the value for k is given as *df* in the `logLik()` output. For our three models, the *df* values are:

- **M1:** 3 *df* ($\hat{\beta}_0$, $\hat{\beta}_1$, RMSE)
- **M2:** 4 *df* ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, RMSE)
- **M3:** 4 *df* ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, RMSE)

This penalty-adjusted value is called Akiake's Information Criteria (AIC). These values can be compared directly, so long as:

- The exact same data is used to fit the models, and
- The exact same outcome is used to fit the models.

Smaller values of the AIC indicate a more likely model.

```
-2 * logLik(lm.1)[1] + 2*3 #Model 1
```

```
## [1] 53.2
```

```
-2 * logLik(lm.2)[1] + 2*4 #Model 2
```

```
## [1] 49.9
```

```
-2 * logLik(lm.3)[1] + 2*4 #Model 3
```

```
## [1] 56.8
```

Arranging these, we find that Model 2 (AIC = 49.9) is the most likely model given the data and the candidate set of models. This is the "best" model given the candidate set of models and the data.

Corrected AIC (AICc): Adjusting for Bias Based on Sample Size and Model Complexity

Although AIC has a penalty correction that should account for some bias, it turns out that when the number of parameters is large relative to the sample size, AIC is still biased in favor of models that have more parameters. This led Hurvich & Tsai (1989) to propose a second-order bias corrected AIC measure (AICc) computed as

$$\text{AIC}_\text{c} = \text{Deviance} + 2(k)\left(\frac{n}{n-k-1}\right)$$

where k is, again, the number of estimated parameters, and n is the sample size used to fit the model. Note that when n is very large (especially relative to k) that the last term is essentially 1 and the AICc value would basically reduce to the AIC value. When n is small relative to k this will add more of a penalty to the deviance. *The recommendation is to pretty much always use AICc rather than AIC when selecting models.*

Below, we will compute the AICc for each of the three candidate models. (Note that we use $n = 52$ cases for the computation for all the models in this data.)

```
n = 52

# Compute AICc for Model 1, B, and C
-2 * logLik(lm.1)[[1]] + 2 * 3 * n / (n - 3 - 1) #Model 1
```

```
## [1] 53.7
```

```
-2 * logLik(lm.2)[[1]] + 2 * 4 * n / (n - 4 - 1) #Model 2
```

```
## [1] 50.7
```

```
-2 * logLik(lm.3)[[1]] + 2 * 4 * n / (n - 4 - 1) #Model 3
```

```
## [1] 57.7
```

Based on the AIC_c values, the best model is again Model 2. It is the most likely model given the data and the six candidate models.

Use AICc() Function

In practice, we will use the `AICc()` function from the **AICcmodavg** package to compute the AICc value directly.

```
AICc(lm.1)
```

```
## [1] 53.7
```

```
AICc(lm.2)
```

```
## [1] 50.7
```

```
AICc(lm.3)
```

```
## [1] 57.7
```

In summary, here are the three candidate models, and their AICc values.

Model	LL	K	AICc
Model 2	-20.9	4	50.7
Model 1	-23.6	3	53.7
Model 3	-24.4	4	57.7

Because the models are proxies for the scientific hypotheses, we can rank order the scientific hypotheses based on the empirical support for each. Of the three scientific hypotheses, H2 (Peer ratings are attributable to the level of interaction faculty members have with students.) has the most empirical support. The second most empirically supported hypothesis is H1. (The rating of a program is attributable to the quality of its students.) Finally, H3 has the least amount of empirical support of the three. (Peer ratings are attributable to the research prestige of the programs.)

ΔAICc

How much more empirical support does H2 have than H1 or H3? We can quantify this by computing the difference in AICc values between the best fitting model and all other models.

Since the minimum AICc value in our candidate models was associated with Model 2, we compute the difference between each model's AICc value and Model 2's AICc value. This is referred to as # ΔAICc.

```
# Compute dalta values
AICc(lm.2) - AICc(lm.1) #Model 1
```

```
## [1] -2.96
```

```
AICc(lm.2) - AICc(lm.2) #Model 2
```

```
## [1] 0
```

```
AICc(lm.2) - AICc(lm.3) #Model 3
```

```
## [1] -6.95
```

Model	LL	K	AICc	ΔAICc
Model 2	-20.9	4	50.7	0.00
Model 1	-23.6	3	53.7	2.96
Model 3	-24.4	4	57.7	6.95

Burnham, Anderson, & Huyvaert (2011, p. 25) give rough guidelines for interpreting ΔAICc values. They suggest that models with ΔAICc values less than 2 are plausible, those in the range of 4–7 have some empirical support, those in the range of 9–11 have relatively little support, and those greater than 13 have essentially no empirical support. Using these criteria:

- Model 2 has a lot of empirical support
- Model 1 and Model 3 both have less empirical support than Model 2, but, nonetheless still both have a fair amount of empirical support

The empirical support is not unequivicably in favor of H2. There is empirical support for all three models. This might mean that we cannot ultimately reject H1 and H3. Or at least that we need to collect additional data to examine the hypotheses.

Relative Likelihood and Evidence Ratios

Onw way we mathematically formalize the strength of evidence for each model is to compute the relative likelihood. To compute the relative likelihood,

$$\text{Relative Likelihood} = e^{\frac{1}{2}(\Delta AICc)}$$

The relative likelihood provides the likelihood of each of the candidate models, given the set of candidate models and the data.

```
exp(-1/2 * 2.96) #Model 1
```

```
## [1] 0.228
```

```
exp(-1/2 * 0.00) #Model 2
```

```
## [1] 1
```

```
exp(-1/2 * 6.95) #Model 3
```

```
## [1] 0.031
```

Model	LL	K	AICc	ΔAICc	Rel. Lik.
Model 2	-20.9	4	50.7	0.00	1.000
Model 1	-23.6	3	53.7	2.96	0.228
Model 3	-24.4	4	57.7	6.95	0.031

These quantities allow evidentiary statements for comparing the scientific hypothese. These are referred to as *evidence ratios*. To compute an evidence ratio, we divide the relative likelihood for any two hypotheses. This will quantify how much more likely one hypothesis is than another given the data. For example, *given the data*,

- The empirical support for Hypothesis H2 is 4.4 (1/.228) times that of the empirical support for Hypothesis H1.
- The empirical support for Hypothesis H2 is 32 (1/.031) times that of the empirical support for Hypothesis H1.

In general, software that computes evidence ratios do so for each model relative to the candidate model with the highest relative likelihood. The resulting evidence ratios allow for a comparison of each hypothesis to the most empirically supported hypothesis. Of course, given the relative likelihood for any two hypotheses, you can always compute the evidence ratio between the associated hypotheses.

- The empirical support for Hypothesis H1 is 7.4 (.228/.031) times that of the empirical support for Hypothesis H3.

Model	LL	K	AICc	ΔAICc	Rel. Lik.	ER
Model 2	-20.9	4	50.7	0.00	1.000	1.00
Model 1	-23.6	3	53.7	2.96	0.228	4.39
Model 3	-24.4	4	57.7	6.95	0.031	32.33

Model Probability

Also referred to as Akaike Weights (w_i), model probabilities provide a numerical measure of the probability of each model given the data and the candidate set. It can be computed as

$$w_i = \frac{\text{Relative Likelihood for Model J}}{\sum_j \text{Relative Likelihood}}$$

```
sum_rel = 1 + .228 + .031

.228 / sum_rel #Model 1
```

```
## [1] 0.181
```

```
1 / sum_rel #Model 2
```

```
## [1] 0.794
```

```
.031 / sum_rel #Model 3
```

```
## [1] 0.0246
```

These values can be interpreted as the probability of the hypothesis given the data and the candidate set of models. For example, given the data and the candidate set of models:

- The probability of Hypothesis H1 is 0.181.
- The probability of Hypothesis H2 is 0.794.
- The probability of Hypothesis H3 is 0.025.

Model	LL	K	AICc	ΔAICc	Rel. Lik.	ER	w_i
Model 2	-20.9	4	50.7	0.00	1.000	1.00	0.795
Model 1	-23.6	3	53.7	2.96	0.228	4.39	0.181
Model 3	-24.4	4	57.7	6.95	0.031	32.33	0.025

Using the `aictab()` Function

We will use the `aictab()` function from the **AICcmodavg** package to compute many of the model evidence values directly from the `lm()` fitted models. This function takes a list of models in the candidate set (it actually has to be an R list). The optional argument `modnames=` is a vector of model names associated with the models in the candidate set.

```
myAIC = aictab(
  cand.set = list(lm.1, lm.2, lm.3),
  modnames = c("Model 1", "Model 2", "Model 3")
  )

# View table
myAIC
```

	Modnames	K	AICc	Delta_AICc	ModelLik	AICcWt	LL	Cum.Wt
2	Model 2	4	50.7	0.00	1.000	0.795	-20.9	0.795
1	Model 1	3	53.7	2.96	0.228	0.181	-23.6	0.975
3	Model 3	4	57.7	6.95	0.031	0.025	-24.4	1.000

Note the output includes the number of parameters (`K`) and AICc value (`AICc`) for each candidate model, and prints them in order from most likely to least likely based on the AICc. It also includes the ΔAICc values and the model probabilities (`AICcWt`) and log-likelihood (`LL`) values. The `Cum.Wt` column gives the cumulative model probabilies. (For example the probability of H2 or H1 is 0.98.)

We have to compute the evidence ratios separately. We do this using the `evidence()` function. This function takes the output from the `aictab()` function as well as the names from that table (given in the `modnames=` argument) for the two models you want to compute the evidence ratio for.

```
# Evidence Ratios
evidence(myAIC, model.high = "Model 2", model.low = "Model 1")
```

```
##
## Evidence ratio between models 'Model 2' and 'Model 1':
## 4.39
```

```
evidence(myAIC, model.high = "Model 2", model.low = "Model 2")
```

```
##
## Evidence ratio between models 'Model 2' and 'Model 2':
## 1
```

```
evidence(myAIC, model.high = "Model 1", model.low = "Model 3")
```

```
##
## Evidence ratio between models 'Model 1' and 'Model 3':
## 7.36
```

Pretty-Printing Model Evidence Tables in Markdown

We can use the `data.frame()` function to coerce the output from the `aictab()` function into a data frame. Then we can use **dplyr** functions to re-order, re-name and add columns to the evidence table. Lastly, we can use `kable()` to format the table for pretty-printing in Markdown.

```
x = data.frame(myAIC) %>%
  select(
    Model = Modnames,
    LL, K, AICc, Delta_AICc,
    w_i = AICcWt
  ) %>%
  mutate(
    ER = max(w_i) / w_i
  )

# Here we employ indexing to change the fifth column name
# We use LaTeX math notation in the names to use the Greek letter Delta
names(x)[5] = '$\\Delta$AICc'

# Here we employ indexing to change the sixth column name
# We use LaTeX math notation to write a subscript
names(x)[6] = '$w_i$'

kable(x, caption = "Table of Model Evidence for Three Candidate Models. (LL = Log-Likelihood; K = Model df; $w_i$ =
```

Table 9: Table of Model Evidence for Three Candidate Models. (LL = Log-Likelihood; K = Model df; w_i = Model Probability; ER = Evidence Ratio)

Model	LL	K	AICc	ΔAICc	w_i	ER
Model 2	-20.9	4	50.7	0.00	0.795	1.00
Model 1	-23.6	3	53.7	2.96	0.181	4.39
Model 3	-24.4	4	57.7	6.95	0.025	32.33

Some Final Thoughts

Based on the different quantifications:

- Hypothesis H2 has the most empirical support.
- There is some empirical support for Hypothesis H1 relative to the other two hypotheses.
- There is very little empirical support for H3 relative to the other two hypotheses.

This might mean that in practice we focus on the the first and second hypotheses, reporting and discussing results from both M1 and M2. We can get a summary of the model rankings along with qualitative descriptors of the empirical support (weight) using the `confset()` function. The `method="ordinal"` argument rank orders the models for us.

```
confset(
  cand.set = list(lm.1, lm.2, lm.3),
  modnames = c("Model 1", "Model 2", "Model 3"),
  method = "ordinal"
  )
```

```
##
## Confidence set for the best model
##
## Method:   ordinal ranking based on delta AIC
##
## Models with substantial weight:
##         K AICc Delta_AICc AICcWt
## Model 2 4 50.7          0   0.79
##
##
## Models with some weight:
##         K AICc Delta_AICc AICcWt
## Model 1 3 53.7       2.96   0.18
## Model 3 4 57.7       6.95   0.02
##
##
## Models with little weight:
##      K AICc Delta_AICc AICcWt
##
##
## Models with no weight:
##      K AICc Delta_AICc AICcWt
```

It is important to note that it is ultimately the set of scientific hypotheses that we are evaluating, using the fit from the associated statistical models to a set of data. If we use a different set of data, we may have a whole new ranking of models, and thus the empirical support is linked to the data.

It is also important to note that we are not evaluating individual predictors in the models, just the model as a whole. Because of this, it is not appropriate to remove preidctors after adopting a model(s).

Lastly, the use of p-values is not compatible with the use of model-level selection methods such as information criteria. See Anderson (2008) for more detail. Because of this, it is typical to not even report p-values when carrying out this type of analysis.

References

Anderson, D. R. (2008). *Model based inference in the life sciences: A primer on evidence*. New York: Springer.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.

Hurvich, C., & Tsai, C.-L. (1989). Regression and time series model selection in small samples, *Biometrika*, *76*, 297–307.