Information Criteria for Model Selection

2020-02-07

Preparation

In this set of notes, you will use information theoretic approaches (e.g., information criteria) to select a "best" model from a set of candidate models (built in the previous set of notes). To do so, we will use the *usnews.csv* dataset (see the data codebook) to examine the factors that underlie the ratings our academic peers give to graduate programs of education.

```
# Load libraries
library(broom)
library(educate) #Need version 0.1.0.1
library(MuMIn)
library(patchwork)
library(tidyverse)

# Import data
usnews = read_csv("~/Documents/github/epsy-8252/data/usnews.csv")

# View data
head(usnews)
```

```
# A tibble: 6 x 13
   rank school score  peer expert_score gre_verbal gre_quant doc_accept
  <dbl> <chr>  <dbl> <dbl>        <dbl>      <dbl>     <dbl>      <dbl>
1     1 Harva~   100   4.4          4.6        163       159        4.5
2     2 Stanf~    99   4.6          4.8        162       160        6.1
3     3 Unive~    96   4.2          4.3        156       152       29.1
4     3 Unive~    96   4.1          4.5        163       157        5
5     3 Unive~    96   4.3          4.5        155       153       26.1
6     6 Johns~    95   4.1          4.1        164       162       27.4
# ... with 5 more variables: phd_student_faculty_ratio <dbl>,
#   phd_granted_per_faculty <dbl>, funded_research <dbl>,
#   funded_research_per_faculty <dbl>, enroll <dbl>
```

Working Hypotheses and Candidate Models

Recall previously that we have three **scientific working hypotheses** about how academics perceive and, ultimatley rate, graduate programs:

- **H1:** Student-related factors drive the perceived academic quality of graduate programs in education.
- **H2:** Faculty-related factors drive the perceived academic quality of graduate programs in education.
- **H3:** Institution-related factors drive the perceived academic quality of graduate programs in education.

These working hypotheses were translated into the following three statistical models:

Model 1

$$\text{Peer Rating}_i = \beta_0 + \beta_1(\text{GREQ}_i) + \beta_2(\text{GREQ}_i^2) + \beta_3(\text{GREQ}_i^3) + \epsilon_i$$

Model 2

$$\text{Peer Rating}_i = \beta_0 + \beta_1(\text{Funded research}_i) + \beta_2(\text{Funded research}_i^2) + \beta_3(\text{PhDs granted}_i) + \beta_4(\text{PhDs granted}_i^2) + \epsilon_i$$

Model 3

$$\text{Peer Rating}_i = \beta_0 + \beta_1(\text{PhD acceptance rate}_i) + \beta_2(\text{PhD student-to-faculty ratio}_i) + \beta_3(\text{Enrollment}_i) + \epsilon_i$$

In these models, peer rating, funded research, Ph.D.s granted, Ph.D. acceptance rate, enrollment, and Ph.D. student-to-faculty ratio were all log-transformed.

```
# Drop rows with missing data
# Create log-transformed variables
educ = usnews %>%
  drop_na() %>%
  mutate(
    Lpeer = log(peer),
    Lfunded_research_per_faculty = log(funded_research_per_faculty),
    Lphd_granted_per_faculty = log(phd_granted_per_faculty + 1),
    Ldoc_accept = log(doc_accept),
    Lphd_student_faculty_ratio = log(phd_student_faculty_ratio + 1),
    Lenroll = log(enroll)
    )

# Fit Model 1
lm.1 = lm(Lpeer ~ 1 + gre_quant + I(gre_quant^2) + I(gre_quant^3), data = educ)

# Fit Model 2
lm.2 = lm(Lpeer ~ 1 + Lfunded_research_per_faculty + I(Lfunded_research_per_faculty^2) + Lphd_granted_per_faculty

# Fit Model 3
lm.3 = lm(Lpeer ~ 1 + Ldoc_accept + Lenroll + Lphd_student_faculty_ratio, data = educ)
```

Log-Likelihood

Recall that the likelihood gives us the probability of a particular model given a set of data and assumptions about the model, and that the log-likelihood is just a mathematically convenient transformation of the likelihood. Log-likelihood values from different models can be compared, so long as:

- The exact same data is used to fit the models,
- The exact same outcome is used to fit the models, and
- The assumptions underlying the likelihood (independence, distributional assumptions) are met.

In all four models we are using the same data set and outcome, and the assumptions seem reasonably tenable for each of the four fitted candidate models. This suggests that the likelihood (or log-likelihood) can provide some evidence as to which of the four candidate models is most probable. Below we compute the log-likelihood values for each of the four candidate models.

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

```
'log Lik.' 95.87649 (df=5)
```

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

```
'log Lik.' 98.24717 (df=6)
```

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

```
'log Lik.' 101.4242 (df=5)
```

Note that the log-likelihood values are also available from the `glance()` function's output (in the `logLik` column).

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

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
1     0.409         0.394 0.112      27.2 1.96e-13     4   95.9 -182. -168.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

These values suggest that the model with the highest probability given the data and set of assumptions is Model 3; 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*. Deviance is a measure of model-data 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] -191.753
```

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

```
[1] -196.4943
```

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

```
[1] -202.8483
```

Here, the model that produces the lowest amount of model-data error is Model 3; it has the lowest deviance value. Since the deviance just multiplies the log-likelihood values by a constant, it produces the same rank ordering of the candidate models. Thus, whether you evaluate using the likelihood, the log-likelihood, or the deviance, you will end up with the same ordering of candidate models. Using deviance, however, has the advantages of having a direct relationship to model error, so it is more interpretable. It is also more closely aligned with other model measures associated with error that we commonly use (e.g., SSE, R^2).

Akiake's Information Criteria (AIC)

Remember that lower values of deviance indicate the model (as defined via the set of parameters) is more likely (lower model-data error) given the data and set of assumptions. However, in practice we cannot directly compare the deviances since the models include a different number of parameters. It was not coincidence that our most probable candidate model also had the highest number of predictors.

To account for this, we will add a penalty term to the deviance based on the number of parameters estimated in the model. This penalty-adjusted value is called Akiake's Information Criteria (AIC).

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

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

- **M1:** 5 *df* ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, RMSE)
- **M2:** 6 *df* ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, RMSE)
- **M3:** 5 *df* ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, RMSE)

Just as with the deviance, smaller AIC values indicate a more likely model.

```
-2 * logLik(lm.1)[1] + 2*5 #Model 1
```

```
[1] -181.753
```

```
-2 * logLik(lm.2)[1] + 2*6 #Model 2
```

```
[1] -184.4943
```

```
-2 * logLik(lm.3)[1] + 2*5 #Model 3
```

```
[1] -192.8483
```

Arranging these, we find that again Model # (AIC = -192.8) is the most likely candidate model given the data and candidate set of models. We can also compute the AIC via the `AIC()` function.

```
# Compute AIC value for Model 1
AIC(lm.1)
```

```
[1] -181.753
```

Lastly, we note that the AIC value is produced as a column in the model-level output. (Note that the `df` column from `glance()` does NOT give the number of model parameters.)

```
# Model-level output for Model 1
glance(lm.1)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
1     0.409         0.394 0.112      27.2 1.96e-13     4   95.9 -182. -168.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

Empirical Support for Hypotheses

Because the models are proxies for the scientific working hypotheses, the AIC ends up being a measure of empirical support for any particular hypothesis—after all, it takes into account the data (empirical evidence) and model complexity. In practice, we can use the AIC to rank order the models, which results in a rank ordering of the scientific working hypotheses based on the empirical support for each. Ranked in order of empirical support, the three scientific working hypotheses are:

- Peer ratings are attributable to institution-related factors. This hypothesis has the most empirical support of the three working hypotheses, given the data and other candidate models.
- Peer ratings are attributable to faculty-related factors.
- Peer ratings are attributable to student-related factors. This hypothesis has the least amount of empirical support of the three working hypotheses, given the data and other candidate models.

It is important to remember that the phrase "given the data and other candidate models" is highly important. Using AIC to rank order the models results in a *relative ranking of the models*. It is not able to rank any hypotheses that you didn't consider as part of the candidate set of scientific working hypotheses. Moreover, the AIC is a direct function of the likelihood which is based on the actual model fitted as a proxy for the scientific working hypothesis. If the predictors used in any of the models had been different, it would lead to different likelihood and AIC values, and potentially a different rank ordering of the hypotheses.

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

Based on the AIC values for the three candidate models we ranked the hypotheses based on the amount of empirical support:

Table 1
Working Hypotheses Rank Ordered by the Amount of Empirical Support as Measured by the AIC

Hypothesis	AIC
Institution-related factors	-192.8
Faculty-related factors	-184.5
Student-related factors	-181.8

Although AIC has a penalty correction that should account for model complexity, 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 the first candidate model. (Note that we use $n = 122$ cases for the computation for all the models in this data.)

```
n = 122
k = 5

# Compute AICc for Model 1
-2 * logLik(lm.1)[[1]] + 2 * k * n / (n - k - 1) #Model 1
```

```
[1] -181.2357
```

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

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

```
[1] -181.2357
```

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

```
[1] -183.7639
```

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

```
[1] -192.3311
```

Based on the AIC_c values, the model with the most empirical support given the data and four candidate models is Model 3. Again, because the models are proxies for the scientific hypotheses, we can rank order the scientific hypotheses based on the empirical support for each.

Table 2
Working Hypotheses Rank Ordered by the Amount of Empirical Support

Hypothesis	AICc
Institution-related factors	-192.3
Faculty-related factors	-183.8
Student-related factors	-181.2

Model-Selection Uncertainty

When we adopt one model over another, we are introducing some degree of selection uncertainty into the scientific process. It would be nice if we can quantify and report this uncertainty, and this is the real advantage of using information criteria for model selection; it allows us to quantify the uncertainty we have when we select any particular candidate model.

The amount of model selection uncertainty we have depends on the amount of empirical support each of the candidate models has. For example, if one particular candidate model has a lot of empirical support and the rest have very little empirical support we would have less model uncertainty than if all of the candidate models had about the same amount of empirical support.

Since we measure the empirical support each hypothesis has by computing the AICc for the associated candidate model, we can look at how much more empirical support the most supported hypothesis has relative to each of the other working hypotheses by computing the difference in AICc values between the best fitting model and each of the other candidate models. This measure is referred to as # ΔAICc.

In our example, the hypothesis with the most empirical support was the institution-related factors model as measured in Model 3.

```
# Compute delta values
AICc(lm.1) - AICc(lm.3) #Student-related factors
```

```
[1] 11.09535
```

```
AICc(lm.2) - AICc(lm.3) #Faculty-related factors
```

```
[1] 8.567177
```

```
AICc(lm.3) - AICc(lm.3) #Institution-related factors
```

```
[1] 0
```

Table 3
Working Hypotheses Rank Ordered by the Amount of Empirical Support

Hypothesis	AICc	ΔAICc
Institution-related factors	-192.3	0.0
Faculty-related factors	-183.8	8.6
Student-related factors	-181.2	11.1

Burnham et al. (2011, p. 25) give rough guidelines for interpreting ΔAICc values. They suggest that hypotheses 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:

- The institution-related factors hypothesis (Model 3) has the most empirical support.
- The faculty-related factor hypothesis (Model 2) is plausible, although it has relatively little empirical support compared to the institution-related factors hypothesis.
- The student-related factor hypothesis (Model 1) has essentially no empirical support relative to the institution-related factor hypothesis.

Relative Likelihood and Evidence Ratios

Onw way we mathematically formalize the strength of evidence for each model is to compute the relative likelihood. The relative likelihood provides the likelihood of each of the candidate models, given the set of candidate models and the data. To compute the relative likelihood,

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

```
# Institution-related factors
exp(-1/2 *  0.00)
```

```
[1] 1
```

```
# Faculty-related factors
exp(-1/2 *  8.6)
```

```
[1] 0.01356856
```

```
# Student-related factors
exp(-1/2 * 11.1)
```

```
[1] 0.003887457
```

Table 4
Working Hypotheses Rank Ordered by the Amount of Empirical Support

Hypothesis	AICc	ΔAICc	Rel. Lik.
Institution-related factors	-192.3	0.0	1.000
Faculty-related factors	-183.8	8.6	0.014
Student-related factors	-181.2	11.1	0.004

Note. Rel. Lik. = Relative Likelihood

1/0.014 = 71.4

0.014/1 = .014

These quantities allow us to compute *evidence ratios*, which are evidentiary statements for comparing any two scientific hypotheses. Evidence ratios quantify how much more empirical support one hypothesis has versus another. To obtain an evidence ratio, we divide the relative likelihood for any two hypotheses. As another example,

- The empirical support for the institution-related factors hypothesis is 71.4 times that of the empirical support for the faculty-related factors hypothesis. (To obtain this we computed $1/.014 = 71.4$.)
- The empirical support for the institution-related factors hypothesis is 250 times that of the empirical support for the student-related factors hypothesis. (To obtain this we computed $1/.004 = 250$.)

Model Probabilities

Also referred to as an Akaike Weight (w_i), a model probability provides a numerical measure of the probability of each model given the data and the candidate set of models. It can be computed as:

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

```
# Compute sum of relative likelihoods
sum_rel = 1.000000000 + 0.01356856 + 0.003887457

# Institution-related factors
1.000000000 / sum_rel
```

```
[1] 0.9828435
```

```
# Faculty-related factors
0.01356856 / sum_rel
```

```
[1] 0.01333577
```

```
# Student-related factors
0.003887457 / sum_rel
```

```
[1] 0.003820762
```

Since the models are proxies for the working hypotheses, the model probabilities can be used to provide probabilities of each working hypothesis as a function of the empirical support. Given the data and the candidate set of working hypotheses:

- The probability of the institution-related factors hypothesis is 0.983.
- The probability of the faculty-related factors hypothesis is 0.013.
- The probability of the student-related factors hypothesis is 0.004.

This suggests that it is highly probably that the institution-related factors hypothesis is the "best" (closest to truth). There is a tiny probability that the faculty-related factors hypothesis is "best"and almost no probability the student-related factors hypothesis is "best".

Table 5
Working Hypotheses Rank Ordered by the Amount of Empirical Support

Hypothesis	AICc	ΔAICc	Rel. Lik.	AICc Weight
Institution-related factors	-192.3	0.0	1.000	0.983
Faculty-related factors	-183.8	8.6	0.014	0.014
Student-related factors	-181.2	11.1	0.004	0.004

Note. Rel. Lik. = Relative Likelihood

Tables of Model Evidence

We will use the `model.sel()` function from the **MuMIn** package to compute and create a table of 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 `rank=` is an optional function that sets the criteria on which to rank the models in the candidate set. (The default if you do not set the `rank=` argument is to use the AICc criterion.)

```
#Create table of model evidence
model_evidence = model.sel(
  object = list(lm.1, lm.2, lm.3),
  rank = "AICc"
  )
```

```
# View output
model_evidence
```

```
Model selection table
    (Int) gre_qnt gre_qnt^2 gre_qnt^3 Lfn_rsr_per_fcl Lfn_rsr_per_fcl^2
3   1.309
2   1.214                                    -0.1508           0.02381
1 779.700  -15.43    0.1017 -0.000223
  Lph_grn_per_fcl Lph_grn_per_fcl^2 Ldc_acc     Lnr Lph_std_fcl_rat df  logLik
3                                  -0.1089 0.01452          0.1287  5 101.424
2          0.2998           -0.1393                                 6  98.247
1                                                                   5  95.876
    AICc delta weight
3 -192.3  0.00  0.983
2 -183.8  8.57  0.014
1 -181.2 11.10  0.004
Models ranked by AICc(x)
```

Note the output includes both the coefficient estimates and the model evidence for each model in the candidate set. The model evidence provided for each model includes the number of parameters (`df`), log-likelihood (`logLik`), AICc value (`AICc`), ΔAICc value (`delta`), and the model probability (`weight`). It also rank orders the models based on the AICc criterion. The models are printed in order from the model with the most empirical evidence (Model 3) to the model with the least amount of empirical evidence (Model 1) based on the AICc.

Pretty Printing Tables of Model Evidence

We can pipe the table of model evidence into the `kable()` function to format the table for pretty-printing in RMarkdown. Here I use **dplyr** functions to select and rename the columns from the `model.sel()` output that correspond to the actual model evidence (omitting the coefficient estimates). One small issue is that the model probabilities have two R classes, so we need to mutate them into numeric values. Finally, we can use functions from the **kableExtra** package (namely `footnote()` and `kable_styling()`) to spruce up the output.

```
# Load libraries for formatting
library(knitr)
library(kableExtra)

# Create data frame to format into table
tab_01 = model_evidence %>%
  mutate(
    Hypothesis = c("H1", "H2", "H3"),
    weight = as.numeric(weight)
  ) %>%
  select(Hypothesis, df, logLik, AICc, delta, weight) %>%
  rename(
    # We can include LaTeX math notation in column names
    # Because \ is a special character we need two \\
    '$K$' = df,
    '$LL$' = logLik,
    '$\\Delta$AICc' = delta,
    'AICc Wt.' = weight
  )
```

```
kable(tab_01,
    format = "latex",
    booktabs = TRUE,
    escape = FALSE,
    caption = "Table of Model Evidence for Three Working Hypotheses",
    digits = 3,
    align = "c",
    row.names = FALSE
    ) %>%
footnote(
  general = "K = Model df; LL = Log-Likelihood; AIC Wt. = Model Probability",
  general_title = "Note.",
  footnote_as_chunk = TRUE
) %>%
kable_styling(latex_options = "HOLD_position")
```

Table 6
Table of Model Evidence for Three Working Hypotheses

Hypothesis	*K*	*LL*	AICc	ΔAICc	AICc Wt.
H1	5	101.424	-192.331	0.000	0.983
H2	6	98.247	-183.764	8.567	0.014
H3	5	95.876	-181.236	11.095	0.004

Note. K = Model df; LL = Log-Likelihood; AIC Wt. = Model Probability

Some Final Thoughts

Based on the model evidence given the data for this candidate set of models:

- The institution-related factors hypothesis has the most empirical support.
- There is very little empirical support for either the faculty-related factor and student-related factors hypotheses relative to the institution-related factors hypothesis.

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

The amount of empirical evidence is also very much relative to the candidate set of models; a different candidate set of models may result in a different rank ordering or interpretation of empirical support. For example, consider if we had not done any exploration of the model's functional form, but instead had just included the linear main-effects for each model.

```r
# Fit models
lm.1_1 = lm(peer ~ 1 + gre_quant + gre_verbal, data = educ)
lm.2_1 = lm(peer ~ 1 + funded_research_per_faculty + phd_granted_per_faculty, data = educ)
lm.3_1 = lm(peer ~ 1 + doc_accept + enroll + phd_student_faculty_ratio, data = educ)

# Compute model evidence
model.sel(
  object = list(lm.1_1, lm.2_1, lm.3_1),
  rank = "AICc"
  )
```

```
Model selection table
   (Int) gre_qnt gre_vrb fnd_rsr_per_fcl phd_grn_per_fcl  doc_acc        enr
3  3.449                                                -0.01128 0.00006721
2  3.026                        0.001338        -0.02629
1 -5.488 0.04677 0.01123
  phd_std_fcl_rat df  logLik  AICc delta weight
3         0.08502  5 -55.495 121.5  0.00   0.55
2                  4 -56.782 121.9  0.40   0.45
1                  4 -68.395 145.1 23.62   0.00
Models ranked by AICc(x)
```

In this example, the rank-ordering of hypotheses ended up being the same, but the evaluation of the empirical support is much different.

- There is about the same amount of empirical support for the institution-related factors hypothesis and the faculty-related factors hypothesis.
- There is still virtually no support for the student-related factors hypothesis.

It is important to note that although information criteria can tell you about the empirical support among a candidate set of models, it cannot say whether that is actually a "good" model. For that you need to look at the assumptions and other measures (e.g., R^2). You still need to do all of the work associated with model-building (e.g., selecting predictors from the substantive literature, exploring functional forms to meet the assumptions).

Statistical Inference and Information Criteria

Finally, it is important to mention that philosophically, information-criteria and statistical inference are two very different ways of measuring statistical evidence. When we use statistical inference for variable selection, the evidence, the p-values, is a measure of how rare an observed statistic (e.g., $\widehat{\beta}_k$, t-value) is under the null hypothesis. The AIC, on the other hand, is a measure of the model-data compatibility accounting for the complexity of the model.

In general, the use of p-values is **not compatible** with the use of information criteria-based model selection methods; see Anderson (2008) for more detail. Because of this, it is typical to not even report p-values when using information criteria for model selection. When using information criteria, however, the standard errors are reported, especially for any "best" model(s). This gives information about the statistical uncertainty that arises because of sampling error.

It is important that you decide how you will be evaluating evidence and making decisions about variable and model selection prior to actually examining the data. Mixing and matching is not cool!

References

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