Information Criteria and Model Selection

Andrew Zieffler
February 26, 2021

Preparation

In this set of notes, you will use information theoretic approaches (e.g., information criteria) to select one (or more) empirically supported model from a set of candidate models. To do so, we will use the *mn-schools.csv* dataset (see the data codebook (http://zief0002.github.io/epsy-8252/codebooks/mn-schools.html)) to examine if (and how) academic "quality" of the student-body (measured by SAT score) is related to institutional graduation rate.

```
# Load libraries
library(AICcmodavg)
library(broom)
library(tidyverse)

# Read in data
mn = read_csv(file = "https://raw.githubusercontent.com/zief0002/epsy-
        8252/master/data/mn-schools.csv")
```

Working Hypotheses and Candidate Models

Over the prior sets of notes, we have considered several models that explain variation in graduation rates. One question we had was which functional form we should adopt (linear, quadratic, log-linear).

$$
\begin{aligned}
\mathbf{Model\ 1:}&\quad \text{Graduation Rate}_i = \beta_0 + \beta_1(\text{SAT}_i) + \epsilon_i \\
\mathbf{Model\ 2:}&\quad \text{Graduation Rate}_i = \beta_0 + \beta_1(\text{SAT}_i) + \beta_2(\text{SAT}_i^2) + \epsilon_i \\
\mathbf{Model\ 3:}&\quad \text{Graduation Rate}_i = \beta_0 + \beta_1\bigg[\ln(\text{SAT}_i)\bigg] + \epsilon_i
\end{aligned}
$$

where all three models have $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$.

```
# Fit candidate models
lm.1 = lm(grad ~ 1 + sat,            data = mn)
lm.2 = lm(grad ~ 1 + sat + I(sat^2), data = mn)
lm.3 = lm(grad ~ 1 + log(sat),       data = mn)
```

Each of these models correspond to a different **scientific working hypothesis** about how graduation rates and median SAT scores are related:

- **H1:** Median SAT scores have a positive relationship with graduation rates that is constant over all levels of SAT.
- **H2:** Median SAT scores have a positive, diminishing relationship with graduation rates for SAT scores below some level. Above this level, median SAT scores have an increasingly negative relationship with graduation rate.
- **H3:** Median SAT scores have a positive, diminsihing relationship with graduation rates.

> These working hypotheses are typically created from the theory and previous empirical work in a substantive area. They need to be translated into a statistical models, which can be quite difficult, especially if there is not a lot of theory to guide this translation.

While the residuals were helpful in ruling out the linear relationship, it was less clear whether the quadratic or log-linear model produced better residuals. We could make that decision on parsimony (the log-linear model is more parsimonious than the quadratic model), but it might be nice to know which model is more empirically supported.

Unfortunately, we cannot use the likelihood ratio test to compare the quadratic and log-linear models, as they are not nested models. Information criteria give us metrics to compare the empirical support for models whether they are nested or not. We will examine several of the more popular information criteria metrics.

Akiake’s Information Criteria (AIC)

The AIC metric is calculated by adding a penalty term to the model's deviance ($-2\ \ln(\text{Likelihood})$).

$$\begin{aligned} AIC &= \text{Deviance} + 2k \\ &= -2\ \ln(\mathcal{L}) + 2k \end{aligned}$$

where k is the number of parameters being estimated in the model. (Recall that the value for k is given as *df* in the `logLik()` output.)

Remember that deviance is similar to error (it measures model-data misfit), and so models that have lower values of deviance are more empirically supported. The problem with deviance, however, is that more complex models tend to have smaller deviances and are, therefore, more supported than their simpler counterparts. Unfortunately, these complex models tend not to generalize as well as simpler models (this is called *overfitting*). The AIC metric's penalty term penalizes the deviance more when a more complex model is fitted; it is offsetting the lower degree of model-data misfit in complex models by increasing the 'misfit' based on the number of parameters.

This penalty-adjusted measure of 'misfit' is called Akiake's Information Criteria (AIC). We compute the AIC for each of the candidate models and the model with the lowest AIC is selected. This model, we say, is the candidate model with the most empirical support. Below we compute the AIC for the model associated with the linear hypothesis.

```
# Compute AIC for linear hypothesis
# logLik(lm.1)
-2*-113.5472 + 2*3
```

```
[1] 233.0944
```

We could also use the `AIC()` function to compute these values directly.

```
# Linear
AIC(lm.1)
```

```
[1] 233.0944
```

```
# Quadratic
AIC(lm.2)
```

```
[1] 227.1166
```

```
# Log-linear
AIC(lm.3)
```

```
[1] 229.5627
```

Based on the AIC values, the candidate model with the most empirical evidence is the quadratic model (Model 2); it has the lowest AIC.

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

```
# Model-level output for linear hypothesis
print(glance(lm.1), width = Inf)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.790         0.783  7.79      117. 4.94e-12     1  -114.  233.  238.
  deviance df.residual  nobs
     <dbl>       <int> <int>
1    1882.          31    33
```

Empirical Support is for the Working Hypotheses

Because the models are proxies for the scientific working hypotheses, the AIC ends up being a measure of empirical support for any particular working hypothesis. Using the AIC, we can rank order the models (and subsequently the working hypotheses) based on their level of empirical support. Ranked in order of empirical support, the three scientific working hypotheses are:

- **H2:** Median SAT scores have a positive, diminishing relationship with graduation rates for SAT scores below some level. Above this level, median SAT scores have an increasingly negative relationship with graduation rate. This hypothesis has the most empirical support of the three working hypotheses, given the data and other candidate models.
- **H3:** Median SAT scores have a positive, diminsihing relationship with graduation rates. This hypothesis has the second most empirical support of the three working hypotheses, given the data and other candidate models.

- **H1:** Median SAT scores have a positive relationship with graduation rates that is constant over all levels of SAT. 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. The ranking of models/working hypotheses is a *relative ranking* of the models’ level of empirical support contingent on the candidate models we included in the comparison and the data we used to compute the AIC.

As such, this method 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.

The ranking of models is also based on the data we have. Even though the quadratic model may not be what we expect substantively, it is more empirically supported than the log-linear model (or the linear model). Why? Well, we only have data in a certain range of median SAT scores. Within this range of scores, the quadratic model is more empirically supported, even after accounting for the additional complexity of the quadratic model. If we had a broader range of median SAT scores, that evidence might support a different model. This is very important. Model 2 is the most empirically supported candidate model GIVEN the three candidate models we compared and the data we used to compute the AIC metric.

> The model selection and ranking is contingent on both the set of candidate models you are evaluating, and the data you have.

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

Models rank-ordered by the amount of empirical support as measured by the AIC.

Hypothesis	k	AIC
Quadratic	3	227.1
Log-linear	4	229.6
Linear	3	233.1

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

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

$$\mathrm{AIC_c} = \mathrm{Deviance} + 2k\left(\frac{n}{n-k-1}\right)$$

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 $n = 33$ cases for the computation for all the models in this data.)

```
n = 33
k = 3

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

```
[1] 233.922
```

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

```
# Linear
AICc(lm.1)
```

```
[1] 233.922
```

```
# Quadratic
AICc(lm.2)
```

```
[1] 228.5452
```

```
# Log-linear
AICc(lm.3)
```

```
[1] 230.3903
```

Based on the AIC_c values, the model with the most empirical support given the data and three candidate models is, again, Model 2. Using these results, we can, again, rank order the models/working hypotheses based on the empirical support for each.

Models rank-ordered by the amount of empirical support as measured by the AICc. The sample size used to fit each model was $n = 33$.

Hypothesis	k	AICc
Quadratic	3	228.5
Log-linear	4	230.4
Linear	3	233.9

Quantifying 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 quantify the empirical support each model/working hypothesis by computing the AICc, we can also quantify 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 model with the mot support 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 quadratic model (Model 2). The ΔAICc values can then be computed for each candidate model by subtracting the AICc for Model 2 from the AICc for the candidate model.

```
# Compute delta AICc value for Model 1
AICc(lm.1) - AICc(lm.2)
```

```
[1] 5.37686
```

```
# Compute delta AICc value for Model 2
AICc(lm.2) - AICc(lm.2)
```

```
[1] 0
```

```
# Compute delta AICc value for Model 3
AICc(lm.3) - AICc(lm.2)
```

```
[1] 1.845128
```

Models rank-ordered by the amount of empirical support as measured by the AICc. The sample size used to fit each model was $n = 33$.

Hypothesis	k	AICc	ΔAICc
Quadratic	3	228.5	0.0
Log-linear	4	230.4	1.8
Linear	3	233.9	5.4

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 quadratic hypothesis (Model 2) is plausible.
- The log-linear hypothesis (Model 3) is also plausible.

- The linear hypothesis (Model 1) is a little less plausible, although it has some empirical support.

This implies that we would have a fair amount of uncertainty actually selecting the quadratic model over the other two models. The empirical evidence has a fair amount of support for all three models, albeit a little less support for the linear hypothesis. This suggests that we adopt more than one model since our empirical evidence can't really differentiate which is 'better.'

Relative Likelihood and Evidence Ratios

One way we can 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,

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

```
# Linear
exp(-1/2 * 5.37686)
```

```
[1] 0.0679876
```

```
# Quadratic
exp(-1/2 * 0)
```

```
[1] 1
```

```
# Log-linear
exp(-1/2 * 1.845128)
```

```
[1] 0.3974985
```

Models rank-ordered by the amount of empirical support as measured by the AICc. The sample size used to fit each model was $n = 33$.

Hypothesis	k	AICc	ΔAICc	Rel($\mathcal{L}$)

Hypothesis	k	AICc	ΔAICc	Rel($\mathcal{L}$)
Quadratic	3	228.5	0.0	1.0
Log-linear	4	230.4	1.8	0.4
Linear	3	233.9	5.4	0.1

Note. Rel($\mathcal{L}$) = Relative Likelihood

The relative likelihood values 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 quadratic hypothesis is 2.52 times that of the empirical support for the log-linear hypothesis. (To obtain this we computed $1/.397 = 2.52$.)
- The empirical support for the quadratic hypothesis is 14.71 times that of the empirical support for the linear hypothesis. (To obtain this we computed $1/.0068 = 14.71$.)

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:

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

```
# Compute sum of relative likelihoods
sum_rel = 0.0679876 + 1 + 0.3974985

# Linear
0.0679876 / sum_rel
```

```
[1] 0.04639252
```

```
# Quadratic
1 / sum_rel
```

```
[1] 0.6823674
```

```
# Log-linear
0.3974985 / sum_rel
```

```
[1] 0.27124
```

Models rank-ordered by the amount of empirical support as measured by the AICc. The sample size used to fit each model was $n = 33$.

Hypothesis	k	AICc	ΔAICc	Rel($\mathcal{L}$)	AICc Weight
Quadratic	3	228.5	0.0	1.0	0.7
Log-linear	4	230.4	1.8	0.4	0.3
Linear	3	233.9	5.4	0.1	0.0

Note. Rel($\mathcal{L}$) = Relative Likelihood

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 quadratic hypothesis is 0.682.
- The probability of the log-linear hypothesis is 0.271.
- The probability of the linear hypothesis is 0.046.

This suggests that the quadratic hypothesis is the most probable. There is also some support for the log-linear hypothesis; it has some non-negligible probability. The linear hypothesis does not have much support; its probability given the candidate models and data is close to zero.

Some Final Thoughts

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

- The quadratic hypothesis has the most empirical support.
- The log-linear hypothesis also has a fair amount of empirical support.
- There is very little empirical support for the linear hypothesis.

Remember, we are computing all of this evidence to select from among the candidate models. Here the empirical evidence is supporting both the quadratic and log-linear model. The data we have does not make it crystal clear about which hypothesis is really more supported.

Recall that the information criteria are a function of the log-likelihood. Log-likelihood values, and thus information criteria, 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 three models we are using the same data set and outcome. However, the assumptions only seem reasonably tenable for the quadratic and log-linear model. That means that we should not really include the linear model in our candidate set of models/working hypotheses. Changing the set of candidate models/working hypotheses can result in different measures of evidence.

Models rank-ordered by the amount of empirical support as measured by the AICc. The sample size used to fit each model was $n = 33$.

Hypothesis	k	AICc	ΔAICc	Rel($\mathcal{L}$)	AICc Weight
Quadratic	3	228.5	0.0	1.0	0.7
Log-linear	4	230.4	1.8	0.4	0.3

Note. Rel($\mathcal{L}$) = Relative Likelihood

Here since we only removed a model from the candidate set, the only thing that changed is the model probabilities (the Akiake weights). Changing the predictors in the candidate models, or adding other models to the candidate set can impact the amount of empirical support and the rank-ordering of hypotheses. If we had a different set of data, we may also have a whole new ranking of models or interpretation of empirical support. The empirical support is linked to the data. The empirical support is very much relative to the candidate set of models and the data we have.

Lastly, 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 evaluating 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., $\hat{\beta}_k$, *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!

Using R to Create a Table of Model Evidence

We will use the `aictab()` function from the `{AICcmodavg}` 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 argument `modnames=` is an optional argument to provide model names that will be used in the output.

```
#Create table of model evidence
model_evidence = aictab(
  cand.set = list(lm.1, lm.2, lm.3),
  modnames = c("Linear", "Quadratic", "Log-linear")
  )

# View output
model_evidence
```

```

Model selection based on AICc:

          K   AICc Delta_AICc AICcWt Cum.Wt      LL
Quadratic  4 228.55       0.00   0.68   0.68 -109.56
Log-linear 3 230.39       1.85   0.27   0.95 -111.78
Linear     3 233.92       5.38   0.05   1.00 -113.55
```

The model evidence provided for each model includes the number of parameters (`K`), AICc value (`AICc`), ΔAICc value (`Delta_AICc`), the Akiake weight/model probability (`AICcWt`), cumulative weight (`Cum.Wt`), and log-likelihood (`LL`). The output also rank orders the models based on the AICc criterion. The models are printed in order from the model with the most empirical evidence (Quadratic) to the working hypothesis with the least amount of empirical evidence (Linear) based on the AICc.

Pretty Printing Tables of Model Evidence for R Markdown Documents

We can format the output from `aictab()` to be used in the `kable()` function. Because there are multiple classes associated with the output from the `aictab()` function, we first pipe `model_evidence` into the `data.frame()` function. Viewing this, we see that the data frame, also includes an additional column that gives the relative likelihoods (`ModelLik`).

```
# Create data frame to format into table
tab_01 = model_evidence %>%
  data.frame()

# View table
tab_01
```

```
    Modnames K     AICc Delta_AICc   ModelLik     AICcWt        LL    Cum.Wt
2  Quadratic 4 228.5452   0.000000 1.00000000 0.68236742 -109.5583 0.6823674
3 Log-linear 3 230.3903   1.845128 0.39749853 0.27124005 -111.7814 0.9536075
1     Linear 3 233.9220   5.376860 0.06798761 0.04639253 -113.5472 1.0000000
```

Then we can use the `select()` function to drop the `LL` and `Cum.Wt` columns from the data frame. The log-likelihood is redundant to the information in the `AICc` column, since AICc is a function of log-likelihood and the other information in the table. The cumulative weight can also easily be computed from the information in the `AICcWt` column.

```
# Drop columns
tab_01 = tab_01 %>%
  select(-LL, -Cum.Wt)

# View table
tab_01
```

```
    Modnames K     AICc Delta_AICc   ModelLik     AICcWt
2  Quadratic 4 228.5452   0.000000 1.00000000 0.68236742
3 Log-linear 3 230.3903   1.845128 0.39749853 0.27124005
1     Linear 3 233.9220   5.376860 0.06798761 0.04639253
```

We can then pipe the `tab_01` data frame into the `kable()` function to format the table for pretty-printing in RMarkdown. I use the `footnote()` and `kable_classic()` functions from the `{kableExtra}` package to add a footnote to the table and also to format the table in our outputted HTML file.

```
# Create knitted table
tab_01 %>%
  kable(
    format = "html",
    booktabs = TRUE,
    escape = FALSE,
    col.names = c("Hypothesis", "k", "AICc", "$\\Delta$AICc",
        "Rel($\\mathcal{L}$)", "AICc Weight"),
    caption = "Models rank-ordered by the amount of empirical support as
        measured by the AICc. The sample size used to fit each model was
        $n=33$.",
    digits = 1,
    table.attr = "style='width:50%;'"
    ) %>%
  footnote(
    general = "Rel($\\mathcal{L}$) = Relative Likelihood",
    general_title = "Note.",
    footnote_as_chunk = TRUE
    ) %>%
  kable_classic()
```

Models rank-ordered by the amount of empirical support as measured by the AICc. The sample size used to fit each model was $n = 33$.

	Hypothesis	k	AICc	ΔAICc	Rel($\mathcal{L}$)	AICc Weight
2	Quadratic	4	228.5	0.0	1.0	0.7

Note. Rel($\mathcal{L}$) = Relative Likelihood

	Hypothesis	k	AICc	ΔAICc	Rel($\mathcal{L}$)	AICc Weight
3	Log-linear	3	230.4	1.8	0.4	0.3
1	Linear	3	233.9	5.4	0.1	0.0

Note. Rel($\mathcal{L}$) = Relative Likelihood

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