Information Criteria

2017-10-22

Preparation

```
# Load libraries
library(tidyverse)
library(lme4)
library(stringr)

# Read in wide data
vocab = read_csv(file = "~/Dropbox/epsy-8282/data/vocab.csv")

# Convert to long data
vocab_long = vocab %>%
  gather(time, score, t8:t11) %>%
  mutate(
    grade = as.integer(str_replace(time, pattern = "t", replacement = "")) - 8
    ) %>%
  arrange(id, grade)
```

Likelihood Estimation in Mixed-Effects Models

With a linear mixed-effects model, the probability distribution we use in the likelihood (or log-likelihood) function is *multivariate normal*. This distribution has a mean of $\boldsymbol{X\beta}$ (since there are more than one distribution we express this as a vector) and a covariance matrix $\boldsymbol{\Phi}$, which identifies the variances and covariances. The density function is,

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\Phi}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\boldsymbol{\Phi})}} \times e^{\left[(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right]}$$

The log-likelihood can be expressed as,

$$\updownarrow = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det\left[\boldsymbol{\Phi}(\boldsymbol{\sigma})\right]\right) - \frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)\boldsymbol{\Phi}^{-1}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)$$

where $\boldsymbol{\sigma}$ is a vector of the random effects, the covariance(s) between the random effects, and the level-1 error variance. The `lmer()` function works to minimize the deviance, which is,

$$\mathcal{D} = n\ln(2\pi) + \ln\left(\det\left[\boldsymbol{\Phi}(\boldsymbol{\sigma})\right]\right) + \left(\boldsymbol{y}-\boldsymbol{X\beta}\right)\boldsymbol{\Phi}^{-1}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)$$

To minimize this function we:

1. Start with initial estimates of the variance and covariance parameters.
2. Use these initial estimates to obtain estimates of the fixed-effects parameters.
3. Use the preliminary fixed-effects estimates to re-estimate the variance and covariance parameters.

4. Continue to iterate, using the new estimates of the varaince and covariance parameters to re-estimate the fixed-effects, which are then used to re-estimate the variances and covariances, etc.

At each iteration, the deviance is computed. The iterations continue until there is little change in the deviance. We can see the results at each iteration by adding the argument `verbose=n` (where `n` is an integer geater than or equal to 1) to the `lmer()` function.

```
lmer.0 = lmer(score ~ 1 + (1 | id), data = vocab_long, REML = FALSE, verbose = 2)
```

```
## npt = 3 , n =  1
## rhobeg =  0.3155298 , rhoend =  3.155298e-07
## start par. =  1.577649 fn =  1012.558
## rho:    0.032 eval:   3 fn:      1009.26 par: 1.26212
## rho:   0.0032 eval:   5 fn:      1009.26 par: 1.26212
## rho:  0.00032 eval:   7 fn:      1009.26 par: 1.27167
## rho:  3.2e-05 eval:   9 fn:      1009.26 par: 1.27151
## rho:  3.2e-06 eval:  11 fn:      1009.26 par: 1.27151
## rho:  3.2e-07 eval:  13 fn:      1009.26 par: 1.27151
## At return
## eval:  16 fn:      1009.2586 par:  1.27151
```

The verbose output here tells us about the optimization over the different iterations. The `fn:` is the value of the function being optimized, in this case, the deviance. The starting deviance was 1012.558. After 16 iterations, the deviance is minimized at 1009.26. The other values here tell us about other parameters being used in the optimization (see `help(minqa::bobyqa)` for more technical information).

Using R to Directly Compute the Likelihood and Log-Likelihood

As we did with `lm()`, we can use the `logLik()` function to directly compute the log-likelihood after we fit a model using the `lmer()` function. From that value, we can also compute the deviance by multiplying the log-likelihood by -2.

```
# Compute log-likelihood
logLik(lmer.0)

# Compute deviance
-504.6293 * -2
```

Akaike Information Criterion (AIC)

Remember that the higher values of log-likelihood (lower values of deviance) indicate the parameters are more likely given the data. You might be thinking that we could use these values to compare models. For example, can we compare the deviances for the model that includes a fixed- and random-effect of intercept and the model that includes fixed- and random-effects of intercept and slope?

```
lmer.1 = lmer(score ~ 1 + grade + (1 + grade | id), data = vocab_long, REML = FALSE)

# Compute log-likelihood
logLik(lmer.1)
```

```
## 'log Lik.' -436.3394 (df=6)
```

```
# Compute deviance
-436.3394 * -2
```

```
## [1] 872.6788
```

The deviance for the second model is smaller than the deviance for the model that only includes intercept effects. Does this imply that the second model fits better? Yes, for these data. However, the two models included a different number of parameters. The second model included more parameters, and like measures such as R^2, we expect that we will improve model-fit by including additional parameters, even if they aren't really important. To account for this, we will add a penalty term to the deviance that penalizes based on the number of parameters each model includes. This penalized deviance is called the AIC.

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

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. The actual AIC values are not really interesting in themselves (they depend on the data, the parameters estimated, and the likelihood function). Information criteria are, however, useful when they are compared to one another (so long as the same data is used and the same outcome). Smaller values of the AIC indicate a more likely model.

For the first model (`lmer.0`),

$$\text{AIC} = 1009.259 + 2(3) = 1015.259$$

For the second model (`lmer.1`),

$$\text{AIC} = 872.6788 + 2(6) = 884.6788$$

Based on the AIC values, we can say that the second model fits better than the first model, even after we penalize for the additional parameters fitted in the second model.

We can also obtain the AIC values directly using the `AIC()` function.

```
AIC(lmer.0)
```

```
## [1] 1015.259
```

```
AIC(lmer.1)
```

```
## [1] 884.6787
```

Bayesian Information Criterion (BIC)

The BIC is another method of penalizing the deviance. BIC is computed as

$$\text{BIC} = \text{Deviance} + k\Big(\ln(n)\Big)$$

where k is, again, the number of estimated parameters, and n is the sample size used to fit the model. In longitudinal data, n is the number of rows used from the long data; in our example, $n = 256$. The penalty for BIC is based on both the number of model parameters and the sample size. We can use the `BIC()` function to compute BIC values.

```
BIC(lmer.0)
```

```
## [1] 1025.894
```

```
BIC(lmer.1)
```

```
## [1] 905.9498
```

AIC or BIC?

The choice of AIC or BIC to select models is somewhat a philosophical decision. TO make this decision, we have to understand that AIC and BIC have different underlying goals:

> "AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model." Penn State Methodology Center

Many times, we do not believe that we have fitted the "correct" model. This has led some researchers to suggest that AIC (which measures "distance" to the truth) might be a better metric for selecting statistical models (e.g., Burnham, Anderson, & Huyvaert, 2011).

Corrected AIC (AICc)

Unfortunately, even though 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). AICc is 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 use the `AICc()` function from the **AICcmodavg** package to compute the AICc values for the two fitted models.

```
# Load library
library(AICcmodavg)

AICc(lmer.0)
```

```
## [1] 1015.354
```

```
AICc(lmer.1)
```

```
## [1] 885.0161
```

Using Information Criterion to Select Models

We can use information criteria to select the "best" model from a set of candidate models. After fitting a set of models we are considering, we then compute the AICc and select the model with the lowest information criterion. We will use the AICc to make this selection.

As an example, consider the three models we considered for the level-1 model:

- Unconditional Means model: $Y_{ij} = \beta_0 + b_{0i} + \epsilon_{ij}$
- Linear effect of grade: $Y_{ij} = \beta_0 + \beta_1(\text{Grade}_{ij}) + b_{0i} + \epsilon_{ij}$
- Quadratic effect of grade: $Y_{ij} = \beta_0 + \beta_1(\text{Grade}_{ij}) + \beta_2(\text{Grade}^2_{ij}) + b_{0i} + \epsilon_{ij}$

```
# Fit candidate models
lmer.0 = lmer(score ~ 1 + (1 | id), data = vocab_long, REML = FALSE)
lmer.1 = lmer(score ~ 1 + grade + (1 | id), data = vocab_long, REML = FALSE)
```

```r
lmer.2 = lmer(score ~ 1 + grade + I(grade^2) + (1 | id), data = vocab_long, REML = FALSE)

# Compute AICc values
AICc(lmer.0)
```

```
## [1] 1015.354
```

```r
AICc(lmer.1)
```

```
## [1] 880.8633
```

```r
AICc(lmer.2)
```

```
## [1] 866.9639
```

Based on the AICc values, we would adopt the model that includes a quadratic effect of grade on vocabulary. That is, given the three candidate models, and the data, the model with the quadratic effect of grade is closest to the unknown TRUE model. When we use information criterion for model selection, there are a few things to keep in mind.

- Philosophically, the use of information criterion are incompatible with the use of p-values. There is no "significantly better"…we select the model with the lowest AICc. Period. Similarly, we do not test the individual predictors in the model we selected for significance.
- The model selected is dependent on the data and the candidate set of models. It may be that had we fitted a different set of candidate models, a different model would have been selected. Similarly, if we had different data, we may have selected a different model (the AICc is, after all, based on the likelihood which is a direct computaton from the data values).
- So long as the same data set and outcome variable are used, AICc values can be compared. This means that models do not have to be nested to make comparisons. Models using completely different sets of fixed- or random-effect structures can be compared to one another.

ΔAICc

Although we don't compute significance to compare the fitted models (e.g., comparing the best and second-best models) it can be quite helpful to determine how much better the "best" model is than the others in the candidate set. One measure of this is to compue the difference between each model's AICc value and the "best" model's AICc value. We call this ΔAICc.

	Model	K	AICc	ΔAICc
3	Quadratic Grade	5	866.96	0.00
2	Linear Grade	4	880.86	13.90
1	Unconditional Means Model	3	1015.35	148.39

Burnham, Anderson, & Huyvaert (2011) 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. In our example,

- The model with the quadratic effect of grade has a lot of empirical support.
- The other two models have essentially no empirical support.

Relative Likelihood

We can also transform the ΔAICc value to a relative likelihood. The relative likelihood provides the likelihood of the candidate model, given the set of candidate models. (Relative likelihoods are also referred to as model

probability weights.) To compute the relative likelihood,

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

For example, the relative likelihood of the unconditional means model is

```
exp(-1/2 * 148.39)
```

```
## [1] 5.991298e-33
```

The unconditional means model is relatively unlikely given the set of candidate models. While the quadratic model has a relative likelihood of

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

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

This model is quite likely given the set of candidate models.

	Model	K	AICc	ΔAICc	Rel. Lik.
3	Quadratic Grade	5	866.96	0.00	1
2	Linear Grade	4	880.86	13.90	0
1	Unconditional Means Model	3	1015.35	148.39	0

Evidence Ratio

By themselves, the relative likelihoods do not give us really any more information than we already had by examining the ΔAICc values. One way to make use of these values is to compute the *evidence ratio* between two models. To do this we compute the ratio of any two model's relative likelihoods.

$$\text{Evidence Ratio} = \frac{\text{Relative Likelihood for Model 1}}{\text{Relative Likelihood for Model 2}}$$

For example, the evidence ratio between the model that includes a linear effect of grade and the unconditional means model is

```
0.000958882847295770964976546757441155933 / 0.000000000000000000000000000000005991417
```

```
## [1] 1.600427e+29
```

This tells us that, given the candidate set of models and the data, the model that includes a linear effect of grade is 160,042,749,035,123,238,169,590,366,208 times more likely than the unconditional means model!!!

It is typical to compare the best model's relative likelihood to each of the other models. If we use the relative likelihood of the best candidate model in the numerator, the resulting evidence ratio allows for a comparison of each candidate model to the best model. Specifically,

$$\text{Evidence Ratio} = \frac{\text{Relative Likelihood for the Best Model}}{\text{Relative Likelihood}}$$

For example, the evidence ratio for the model that includes a linear effect of grade is

```
1.000000000000000000000000000000000000000 / 0.000958882847295770964976546757441155933
```

```
## [1] 1042.88
```

We can interpret this as: *Given the data, and the set of candidate models, the empirical support for the model that includes the quadratic effect of grade is 1043 times that for the model that includes the linear effect of grade.*

For the unconditional means model the evidence ratio is:

```
1.000000000000000000000000000000000000 / 0.000000000000000000000000000000005991417
```

```
## [1] 1.669054e+32
```

We can interpret this as: *Given the data, and the set of candidate models, the empirical support for the model that includes the quadratic effect of grade is 166,905,424,876,953,139,107,289,515,425,792 times that for the unconditional means model.*

Model Probability

Also referred to as Akaike Weights, the model probability provides a numerical measure of the probability of a model given the data and the candidate set. We compute it as

$$\text{Model Probability} = \frac{\text{Relative Likelihood}}{\sum \text{Relative Likelihoods}}$$

For example, the model probabiity of the model that includes a quadratic effect of grade is

```
1.000000000000000000000000000000000000 / (1.000000000000000000000000000000000000 + 0.000958882847295770964976540
```

```
## [1] 0.999042
```

Given the data, and the set of candidate models, the probability of the model that includes a quadratic effect of grade is 0.999.

	Model	K	AICc	ΔAICc	Rel. Lik.	Model Prob.
3	Quadratic Grade	5	866.964	0.000	1.000	0.999
2	Linear Grade	4	880.863	13.899	0.001	0.001
1	Unconditional Means Model	3	1015.354	148.390	0.000	0.000

Direct Computation of the Model Information

We will use the `aictab()` function from the **AICcmodavg** package to compute these values directly from the `lmer()` 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.

```
library(AICcmodavg)

# AICc Table for Model Selection
myAIC = aictab(
  cand.set = list(lmer.0, lmer.1, lmer.2),
  modnames = c("Unconditional means model", "Linear effect of grade", "Quadratic effect of grade")
  )

# View table
myAIC
```

```
##
## Model selection based on AICc:
##
```

```
##                           K    AICc Delta_AICc AICcWt Cum.Wt      LL
## Quadratic effect of grade 5  866.96       0.00      1      1 -428.36
## Linear effect of grade    4  880.86      13.90      0      1 -436.35
## Unconditional means model 3 1015.35     148.39      0      1 -504.63
```

Note the output includes the number of parameters (k) and AICc value for each candidate model, and prints them in order from most likely to least likely based on the AICc. It also includes the ΔAICc value, model probability (AICcWt), cumulative probability (Cum.Wt) and log-likelihood (LL). (If you want the evidence ratios, you have to compute those yourself.)

The output of the aictable (`myAIC`) can be coerced into a data frame, using the `data.frame()` function. We can then add columns, use indexing, submit results to `ggplot()`, etc.

```
data.frame(myAIC)
```

```
##                     Modnames K      AICc Delta_AICc     ModelLik
## 3 Quadratic effect of grade 5  866.9639    0.00000 1.000000e+00
## 2    Linear effect of grade 4  880.8633   13.89948 9.588828e-04
## 1 Unconditional means model 3 1015.3538  148.38996 5.991417e-33
##         AICcWt        LL   Cum.Wt
## 3 9.990420e-01 -428.3619 0.999042
## 2 9.579643e-04 -436.3520 1.000000
## 1 5.985677e-33 -504.6293 1.000000
```

Here I select the first six columns of the data frame, and use the `kable()` function from the **knitr** package to pretty print the table (change the column names, set the number of decimal places.)

```
# View table
knitr::kable(
  data.frame(myAIC)[ , 1:6],
  col.names = c("Model", "K", "AICc", "$\\Delta$AICc", "Rel. Lik.", "Model Prob."),
  digits = 3
  )
```

	Model	K	AICc	ΔAICc	Rel. Lik.	Model Prob.
3	Quadratic effect of grade	5	866.964	0.000	1.000	0.999
2	Linear effect of grade	4	880.863	13.899	0.001	0.001
1	Unconditional means model	3	1015.354	148.390	0.000	0.000

All of this evidence points to the model that includes the quadratic effect of grade as the best candidate model given the data. The model that includes a linear effect of grade and the unconditional means model have almost no empirical support.

Selecting Random Effects

Once we have adopted the functional form for the level-1 (intra-subject) model, we can think about the random effects structure. The adopted level-1 model was

$$\text{Vocabulary}_{ij} = \beta_0^* + \beta_1^*(\text{Grade}_{ij}) + \beta_2^*(\text{Grade}_{ij}^2) + \epsilon_{ij}$$

Here are three possible structures for the random effects.

Random effect of intercept

$$\beta_0^* = \beta_{00} + b_{0i}$$
$$\beta_1^* = \beta_{10}$$
$$\beta_2^* = \beta_{20}$$

Random effect of intercept and linear slope

$$\beta_0^* = \beta_{00} + b_{0i}$$
$$\beta_1^* = \beta_{10} + b_{1i}$$
$$\beta_2^* = \beta_{20}$$

Random effect of intercept, linear slope, and quadratic slope

$$\beta_0^* = \beta_{00} + b_{0i}$$
$$\beta_1^* = \beta_{10} + b_{1i}$$
$$\beta_2^* = \beta_{20} + b_{2i}$$

We can fit each of these models and evaluate the model information.

```
# Fit the three RE structures
lmer.2.0 = lmer(score ~ 1 + grade + I(grade^2) + (1 | id), data = vocab_long, REML = FALSE)
lmer.2.1 = lmer(score ~ 1 + grade + I(grade^2) + (1 + grade | id), data = vocab_long, REML = FALSE)
lmer.2.2 = lmer(score ~ 1 + grade + I(grade^2) + (1 + grade + I(grade^2) | id), data = vocab_long, REML = FALSE)

# Evaluate model information
myAIC = aictab(
  cand.set = list(lmer.2.0, lmer.2.1, lmer.2.2),
  modnames = c("Intercept RE", "Intercept and Linear RE", "Intercept, Linear, and Quadratic RE")
  )

# View table
myAIC
```

```
##
## Model selection based on AICc:
##
##                                    K   AICc Delta_AICc AICcWt Cum.Wt
## Intercept RE                       5 866.96       0.00   0.86   0.86
## Intercept and Linear RE            7 871.15       4.18   0.11   0.96
## Intercept, Linear, and Quadratic RE 10 873.17       6.21   0.04   1.00
##                                        LL
## Intercept RE                       -428.36
## Intercept and Linear RE            -428.35
## Intercept, Linear, and Quadratic RE -426.14
```

The model that includes a random effect of intercept has the most empirical support. Given the candidate set of models and the data, this model has a probability of 0.86. There is also some empirical support for the model that includes random effects for the intercept and slope. However, the support for this model is less than for the model that only includes a random effect for the intercept. The evidence ratio between these models suggests that the model that only includes a random effect for the intercept has 8-times the empirical support than the model that includes a random effect of intercept and slope.

```
# Evidence ratio
exp(-0.5*0) / exp(-0.5*4.18)
```

```
## [1] 8.084915
```

Including Covariates

Once the functional form f the level-1 model and the random effects structure have been adopted, we can evaluate any potential covariates. Here we have a sex covariate (`female`) that we can include. Since there ws some support for the RE structure of the linear term, we will fit models that include the female covariate for the model that include only the RE of intercept and that which includes a RE for both intercept and slope. We will also fit the models without the female covariates for comparison.

```
# Fit the models w/o female
lmer.2.0 = lmer(score ~ 1 + grade + I(grade^2) + (1 | id), data = vocab_long, REML = FALSE)
lmer.2.1 = lmer(score ~ 1 + grade + I(grade^2) + (1 + grade | id), data = vocab_long, REML = FALSE)

# Fit the models w/female
lmer.2.0.f = lmer(score ~ 1 + grade + I(grade^2) + female + (1 | id), data = vocab_long, REML = FALSE)
lmer.2.1.f = lmer(score ~ 1 + grade + I(grade^2) + female + (1 + grade | id), data = vocab_long, REML = FALSE)

# Evaluate model information
myAIC = aictab(
  cand.set = list(lmer.2.0, lmer.2.1, lmer.2.0.f, lmer.2.1.f),
  modnames = c("Intercept RE", "Intercept and Linear RE", "Intercept RE + Female", "Intercept and Linear RE + Female")
  )

# View table
myAIC
```

```
##
## Model selection based on AICc:
##
##                                 K   AICc Delta_AICc AICcWt Cum.Wt      LL
## Intercept RE                    5 866.96       0.00   0.65   0.65 -428.36
## Intercept RE + Female           6 869.00       2.04   0.24   0.89 -428.33
## Intercept and Linear RE         7 871.15       4.18   0.08   0.97 -428.35
## Intercept and Linear RE + Female 8 873.22       6.25   0.03   1.00 -428.32
```

Here the empirical evidence points toward the model that only includes a random effect of intercept and no other covariates. This model is 2.77 times as likely as the model that includes female as a covariate (and only has a random effect of intercept), given the candidate set of models and the data.

```
# Evidence ratio
exp(-0.5*0) / exp(-0.5*2.04)
```

```
## [1] 2.773195
```

In practice, these results (in conjunction with the previous results) would probably suggest that the RE for linear slope could be dropped, and only the RE for intercept is necessary. I might report the both models (with and without female covariate) for the intercept-only RE. WHile the support for not including the female covariate is higher, it is not overwhelming.

References

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.