Model Evaluation and Information Criteria

Felipe José Bravo Márquez

September 27, 2021

Model Evaluation and Information Criteria

- In the context of scientific models, there are two fundamental kinds of statistical error [McElreath, 2020]:
  - **Overfitting**, which leads to poor prediction by learning too much from the data.
  - **Underfitting**, which leads to poor prediction by learning too little from the data.
- There are two common families of approaches to tackle these problems.
  - **Regularization**: a mechanism to tell our models not to get too excited by the data.
  - **Information criteria**: a scoring device to estimate predictive accuracy of our models.
- In order to introduce information criteria, this class must also introduce **information theory**.

The problem with parameters

- In the class of linear regression we learned that including more attributes can lead to a more accurate model.
- However, we have also learned that adding more variables almost always improves the fit of the model to the data, as measured by the coefficient of determination R^2.
- This is true even when the variables you add to a model are just random numbers, with no relation to the outcome.
- So it's no good to choose among models using only fit to the data.

The problem with parameters

- While more complex models fit the data better, they often predict new data worse.
- This means that a complex model will be very sensitive to the exact sample used to fit it.
- This will lead to potentially large mistakes when future data is not exactly like the past data.
- But simple models, with too few parameters, tend instead to underfit, systematically over-predicting or under-predicting the data.
- Regardless of how well future data resemble past data.
- So we can't always favor either simple models or complex models.
- Let's examine both of these issues in the context of a simple data example.

The problem with parameters

- We are going to create a data.frame containing average brain volumes and body masses for seven hominin species.

```
sppnames <- c( "afarensis","africanus","habilis",
               "boisei", "rudolfensis","ergaster",
               "sapiens")
brainvolcc <- c( 438 , 452 , 612, 521, 752, 871,
                 1350 )
masskg <- c( 37.0 , 35.5 , 34.5 , 41.5 , 55.5 ,
             61.0 , 53.5 )
d <- data.frame( species=sppnames , brain=brainvolcc,
                 mass=masskg )
```

- It's not unusual for data like this to be highly correlated.
- Brain size is correlated with body size, across species.
- We will model brain size as a linear function of body size.
- We will fit a series of increasingly complex model families and see which function fits the data best.

Information theory and model performance

- Blabla

Regularization

- Blabla

Information criteria

- Blabla

Using information criteria

- Blabla

Conclusions

- Blabla

References I

McElreath, R. (2020).
Statistical rethinking: A Bayesian course with examples in R and Stan.
CRC press.