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.

The problem with parameters

- 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.
- Each of these models will just be a polynomial of higher degree.

```
reg.ev.1 <- lm( brain ~ mass , data=d )
reg.ev.2 <- lm( brain ~ mass + I(mass^2)
                , data=d )
reg.ev.3 <- lm( brain ~ mass + I(mass^2)
                + I(mass^3),data=d )
reg.ev.4 <- lm( brain ~ mass + I(mass^2)
                + I(mass^3) + I(mass^4),data=d )
reg.ev.5 <- lm( brain ~ mass + I(mass^2)
                + I(mass^3) + I(mass^4)
                + I(mass^5),data=d )
reg.ev.6 <- lm( brain ~ mass + I(mass^2)
                + I(mass^3) + I(mass^4)+
                  I(mass^5)+ I(mass^6),data=d )
```

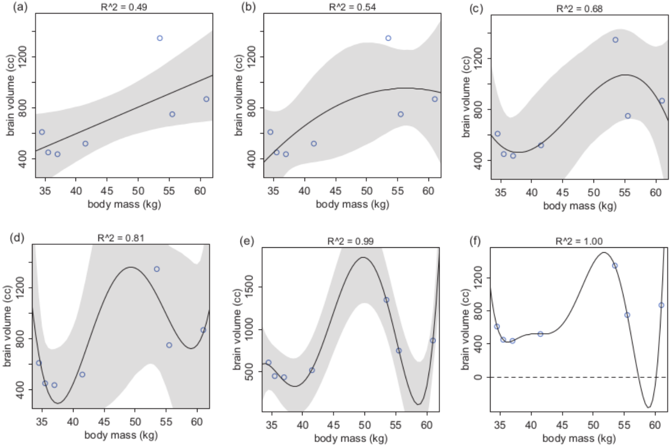
The problem with parameters

- Let's calculate R^2 for each of these models:

```
> summary(reg.ev.1)$r.squared
[1] 0.490158
> summary(reg.ev.2)$r.squared
[1] 0.5359967
> summary(reg.ev.3)$r.squared
[1] 0.6797736
> summary(reg.ev.4)$r.squared
[1] 0.8144339
> summary(reg.ev.5)$r.squared
[1] 0.988854
> summary(reg.ev.6)$r.squared
[1] 1
```

- As the degree of the polynomial defining the mean increases, the fit always improves.
- The sixth-degree polynomial actually has a perfect fit, $R^2 = 1$.

The problem with parameters



Polynomial linear models of increasing degree, fit to the hominin data. Each plot shows the predicted mean in black, with 89% interval of the mean shaded. R^2, is displayed above each plot. (a) First-degree polynomial. (b) Second-degree. (c) Third-degree. (d) Fourth-degree. (e) Fifth-degree. (f) Sixth-degree. Source: [McElreath, 2020].

The problem with parameters

- We can see from looking at the paths of the predicted means that the higher-degree polynomials are increasingly absurd.
- For example panel (f) shows the most complex model, reg.ev.6.
- The fit is perfect, but the model is ridiculous.
- Notice that there is a gap in the body mass data, because there are no fossil hominins with body mass between 55 kg and about 60 kg.
- In this region, the predicted mean brain size from the high-degree polynomial models has nothing to predict, and so the models pay no price for swinging around wildly in this interval.
- The swing is so extreme that at around 58 kg, the model predicts a negative brain size!
- The model pays no price (yet) for this absurdity, because there are no cases in the data with body mass near 58 kg.

The problem with parameters

- Why does the sixth-degree polynomial fit perfectly?
- Because it has enough parameters to assign one to each point of data.
- The model's equation for the mean has 7 parameters:

  $$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + +\beta_5 x_i^5 + \beta_6 x_i^6 + \epsilon_i \quad \forall i$$

  and there are 7 species to predict brain sizes for.
- So effectively, this model assigns a unique parameter to reiterate each observed brain size.
- This is a general phenomenon: If you adopt a model family with enough parameters, you can fit the data exactly.
- But such a model will make rather absurd predictions for yet-to-be-observed cases.

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.