Model Evaluation and Information Criteria

Felipe José Bravo Márquez

September 28, 2021

Model Evaluation and Information Criteria

- In this class we will introduce various concepts for evaluating statistical models.
- According to [McElreath, 2020], there are two fundamental kinds of statistical error:
  - **Overfitting**: models that learn too much from the data leading to poor prediction.
  - **Underfitting**: moderls that learn too little from the data, which also leads to poor prediction.
- We will study 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 some concepts of **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 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 we add to a model have 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.

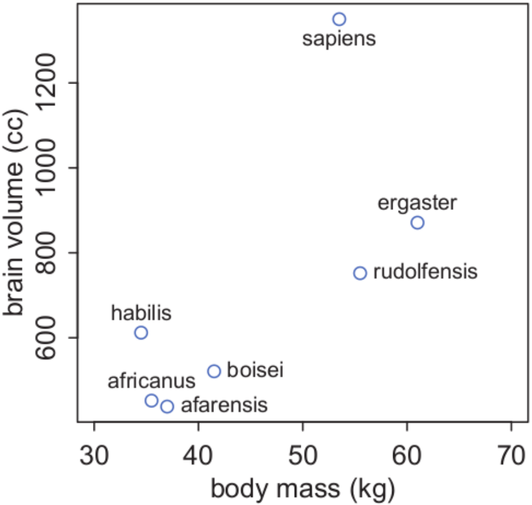
The problem with parameters

- 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



The problem with parameters

- We will model brain size as a 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 )
```

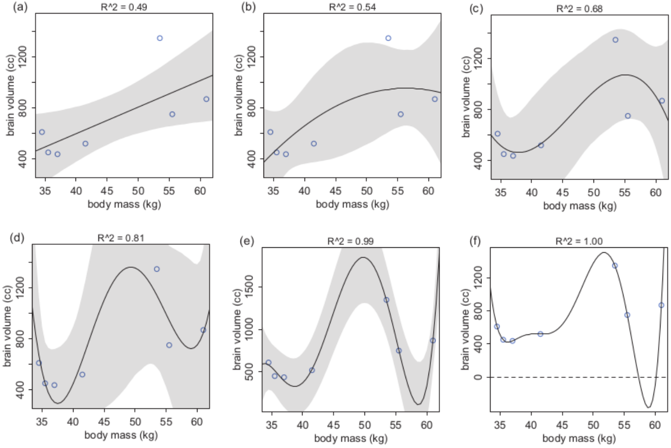
The problem with parameters

- 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, **reg.ev.6** the most complex model makes a perfect fit, 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 models has nothing to predict, so it pays 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.

Too few parameters hurts, too

- The overfit polynomial models manage to fit the data extremely well.
- But they suffer for this within-sample accuracy by making nonsensical out-of-sample predictions.
- In contrast, underfitting produces models that are inaccurate both within and out of sample.
- For example, consider this model of brain volume:

$$y_i = \beta_0 + \epsilon_i \quad \forall i$$

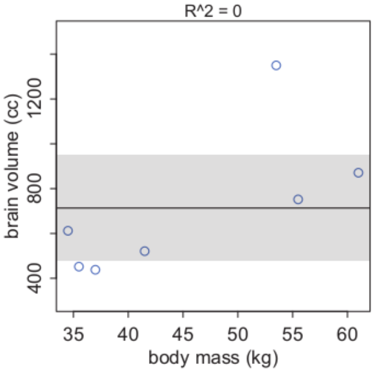
- There are no predictor variables here, just the intercept β_0.
- We can fit this model as follows:

```
> reg.ev.0 <- lm( brain ~ 1 , data=d )
> summary(reg.ev.0)$r.squared
[1] 0
```

- The value of R^2 is 0.

Too few parameters hurts, too

- This model estimates the mean brain volume, ignoring body mass.



- As a result, the regression line is perfectly horizontal and poorly fits both smaller and larger brain volumes.
- Such a model not only fails to describe the sample.
- It would also do a poor job for new data

Roadmap

- The first thing we need to navigate between overfitting and underfitting problems is a criterion of model performance.
- We will see how information theory provides a useful criterion for model evaluation: the **out-of-sample deviance**.
- Once we learn about this cretierion we will se how both regularization and information criteria help to improve and estimate the out-of-sample deviance of a model.
- As usual, we will introduce these concepts with an example.

The Weatherperson

- Suppose in a certain city, a certain weatherperson issues uncertain predictions for rain or shine on each day of the year.
- The predictions are in the form of probabilities of rain.
- The currently employed weatherperson predicted these chances of rain over a 10-day sequence:

Day	1	2	3	4	5	6	7	8	9	10
Prediction	1	1	1	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Observed	🌧	🌧	🌧	☼	☼	☼	☼	☼	☼	☼

- A newcomer rolls into town, and this newcomer boasts that he can best the current weatherperson, by always predicting sunshine.
- Over the same 10 day period, the newcomer's record would be:

Day	1	2	3	4	5	6	7	8	9	10
Prediction	0	0	0	0	0	0	0	0	0	0
Observed	🌧	🌧	🌧	☼	☼	☼	☼	☼	☼	☼

The Weatherperson

- Define hit rate as the average chance of a correct prediction.
- So for the current weatherperson, she gets $3 \times 1 + 7 \times 0.4 = 5.8$ hits in 10 days, for a rate of $5.8/10 = 0.58$ correct predictions per day.
- In contrast, the newcomer gets $3 \times 0 + 7 \times 1 = 7$, for $7/10 = 0.7$ hits per day.
- The newcomer wins.
- Let's compare now the two predictions using another metric, the joint likelihood: $\prod f(y_i; \theta)$ for a frequentist model or $\prod f(y_i|\theta)$ for a Bayesian one.
- The joint likelihood corresponds to the joint probability of correctly predicting the observed sequence.
- To calculate it we must first compute the probability of a correct prediction for each day.
- Then multiply all of these probabilities together to get the joint probability of correctly predicting the observed sequence.

The Weatherperson

- The probability for the current weather person is $1^3 \times 0.4^7 \approx 0.005$.
- For the newcomer, it's $0^3 \times 1^7 = 0$.
- So the newcomer has zero probability of getting the sequence correct.
- This is because the newcomer's predictions never expect rain.
- So even though the newcomer has a high average probability of being correct (hit rate), he has a terrible joint probability (likelihood) of being correct.
- And the joint likelihood is the measure we want.
- Because it is the unique measure that correctly counts up the relative number of ways each event (sequence of rain and shine) could happen.
- In the statistics literature, this measure is sometimes called the **log scoring rule**, because typically we compute the logarithm of the joint probability and report that.

Information and uncertainty

- So we want to use the log probability of the data to score the accuracy of competing models.
- The next problem is how to measure distance from perfect prediction.
- A perfect prediction would just report the true probabilities of rain on each day.
- So when either weatherperson provides a prediction that differs from the target, we can measure the distance of the prediction from the target.

Information and uncertainty

- What kind of distance should we adopt?
- Getting to the answer depends upon appreciating what an accuracy metric needs to do.
- It should appreciate that some targets are just easier to hit than other targets.
- For example, suppose we extend the weather forecast into the winter. Now there are three types of days: rain, sun, and snow.
- Now there are three ways to be wrong, instead of just two.
- This has to be reflected in any reasonable measure of distance from the target, because by adding another type of event, the target has gotten harder to hit.
- Before presenting a distance metric that satisfies the properties described above, we must introduce some concepts from information theory.

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.