COMP6053 lecture: Akaike's information criterion; model reduction

jn2@ecs.soton.ac.uk
Occam's razor

- William of Occam, 1288-1348.
- All else being equal, the simplest explanation is the best one.
Occam's razor

- In statistics, this means a model with fewer parameters is to be preferred to one with more.

- Of course, this needs to be weighed against the ability of the model to actually predict anything...
Why reduce models?

- In keeping with Occam's razor, the idea is to trim complicated multi-variable models down to a reasonable size.

- This is most obvious when we look at multiple regression.

- Do I need all 12 of these predictors? Would a model with only 6 predictors be almost as accurate and thus preferable?
Why reduce models?

- The same logic applies to even the simplest statistical tests.

- A two-sample t-test asks whether the model that says *the two samples come from populations with different means* could be pared down to the simpler model that says *they come from a single population with a common mean.*
Over-fitting

- How greedy can we get?
- In other words, how many predictors, or degrees of freedom, can a model reasonably have?
- A useful absolute ceiling to think about is a model with N-1 binary categorical predictor variables, where N is the sample size.
Over-fitting

![Graph showing over-fitting](image)
Over-fitting

- N-1 predictors would be enough to assign a unique value to each case.
- That would allow the model to explain all variance in the data, but it's clearly an absurd model.
- We "explain" the data by just reading it back to ourselves in full. No compression or explanation is achieved.
Under-fitting

• Thus we want to have fewer predictors in our model than there are cases in the data: usually a lot fewer.

• How minimal can we get?

• The minimal model is to simply explain all the variation in our outcome measure by specifying its mean.
Under-fitting
Under-fitting

- If you had no good model of height, and I asked you the height of the next person you see, your best response is 1.67m (i.e., the UK average).

- All of the variance in the outcome measure remains unexplained, but at least we can say our one-parameter model is economical!
A compromise

- We want a model that is as simple as possible, but no simpler.
- A reasonable amount of explanatory power traded off against model size (number of predictors).
- How do we measure that?
The old way of doing it

- People used to do model reduction through a series of F-tests asking whether a model with one extra predictor explained significantly more of the variance in the dependent or outcome variable.

- This was called "stepwise model reduction", and was done either by pruning from a full model (backwards) or building up from a null model (forwards).
The old way of doing it

● It wasn't a bad way to do it, but one problem was that the model you ended up with could be different depending on the order in which you examined candidate variables.

● To get to a better method we have to look briefly at information theory...
Kullback-Leibler divergence

- Roughly speaking this is a measure of the informational distance between two probability distributions.

- The K-L distance between a real-world distribution and a model distribution tells us how much information is lost by summarizing the phenomenon with that model.

- Minimizing the K-L distance is a good plan.
Maximum likelihood estimation

- A likelihood function gives the probability of observing the data given a certain set of model parameters.

- It's not the same as the probability of a model being true. It's just a measure of how strange the data would be given a particular model.
Coin tossing example

- We throw a coin three times: it comes up heads twice and tails once.

- We have two competing theories about the nature of the coin:
  - A: it's a fair coin, \( p(\text{heads}) = 0.5 \)
  - B: it's a biased coin, \( p(\text{heads}) = 0.8 \).

- There are 3 distinct ways to get 2 heads and 1 tail in 3 throws: HHT, HTH, THH.
Coin tossing example

- Under model A, each of those possibilities has $p = 0.125$, and the total probability of getting two heads and one tail (i.e., the data) is 0.375.

- Under model B, each of those possibilities has $p = 0.128$, and the total probability of two heads, one tail is 0.384.

- The likelihood function is maximized by model B in this case.
Coin tossing example

- We would therefore prefer model B (narrowly) to model A, because it's the model that renders the observed data "less surprising".

- Note that in this case models A and B have the same number of parameters, so there's nothing between them on simplicity, only accuracy.
"Akaike's information criterion"

- In the 1970s he used information theory to build a numerical equivalent of Occam's razor.
Akaike's information criterion

- The idea is that if we knew the true distribution $F$, and we had two models $G_1$ and $G_2$, we could figure out which model we preferred by noting which had a lower K-L distance from $F$.

- We don't know $F$ in real cases, but we can estimate $F-G_1$ and $F-G_2$ from our data.
Akaike's information criterion

- That's what AIC is.
- The model with the lowest AIC value is the preferred one.
- The formula is remarkably simple: 
  \[ AIC = 2K - 2\log(L) \]
  ... where K is the number of predictors and L is the maximized likelihood value.
Akaike's information criterion

- The "2K" part of the formula is effectively a penalty for including extra predictors in the model.

- The "-2 log(L)" part rewards the fit between the model and the data.

- Likelihood values in real cases will be very small probabilities. So "-2 log(L)" will be a large positive number.
How do I use this in R?

- AIC is spectacularly easy to use in R.

- The command is
  \[
  \text{AIC(}\text{model1, model2, model3, ...})
  \]

- This lists the AIC values for all the named models; simply pick the lowest.

- \text{drop1(model)} is also very useful. It gives the AIC value for the models reached by dropping each predictor in turn from this one.
Additional resources

- At this point in the lecture we switch to an interactive R session to show how to use AIC in practice to reduce a model.
- The example data set is the Oscars data from the previous lecture on logistic regression.
- Logistic regression example, but AIC also works with linear regression and any model where a maximum likelihood estimate exists.
Additional materials

- The Python code for generating graphs and the fictional data set used here; also the Python code for generating the fictional Oscars data set.

- The fictional Oscars data set as a text file.

- An R script for analyzing the fictional data set.
Additional materials

- If you want to reproduce the R session used in the lecture, load the above data file and R script into your working directory, and then type this command:

```r
source("aicScript.txt", echo=TRUE)
```