Model Selection I: Why do we select models?

Bob O'Hara

Contents

This Week: Model selection

This week you will:

- find out why model selection is needed
- be able to use AIC and BIC to compare models
- be able to compare hypotheses with F tests

The material is split into 3 modules (because one was getting too long). This is the first, the second is about hypothesis testing, and the third is about exploratory model selection. You should try to follow them in order.

Here we will cover one of the more contentious areas of statistics - how to decide what model to use. One reason for the controversy is the link to hypothesis testing. In essence, hypothesis testing is model selection: each hypothesis (null and alternative - H_0 and H_1) is represented by a model, and hypothesis testing asks whether H_0 can be supported by the data. As we will see, this makes the test depend on the amount of data. As H_0 is usually wrong, this means we are asking if we have enough data to see if H_0 is wrong.

This is not to argue that hypothesis testing is always a bad idea, rather it is one statistical tool, and should (like all tools) be used with caution.

First, a video click here or watch:

Why select models?

It is worth starting out by looking at why we want to select models. It is probably clear when we have specific hypotheses, e.g. when we are asking "does adding cow manure to fields improve yield?" we are asking whether the manure effect is non-zero. So comparing models when it is zero or non-zero is sensible. But when we are asking "which one of these 1000 loci explains how wide people's faces are", one might argue that it is better just to use all of the effects. But using all of the variables comes at a price...

Let us see what happens. We want you to simulate some data, with 100 points and (up to) 90 explanatory variables. The first variable will explain about ~1% of data, the rest will be simulated independently.

```
set.seed(25)
# This first line sets a seed to make sure you all get
```

```r
# the same data
NData <- 100
NExplanatory <- 90
# Create data of explanatory variables
x <- matrix(rnorm(NData*NExplanatory), nrow=NData)
# Calculate "true" mean
mu <- 0.1*x[,1] # true R^2 = 0.1^2/(0.1^2 + 1) = 1%
# Simulate the data
y <- rnorm(NData, mu, 1)
```

Now we want to fit a regression model to the data. We know the true model is $y_i \sim N(0.1x_{1i}, 1)$, but what happens when we add more variables?

We will use a function, `GetR2()`, to fit a model, and return the R^2, and the estimate of the effect of the first variable. We use it multiple times, looping over the vector `1:Nexplanatory`, each time using that number of covariates. Rather than writing a for loop, we use the `sapply()` function, which does the looping internally[1]:

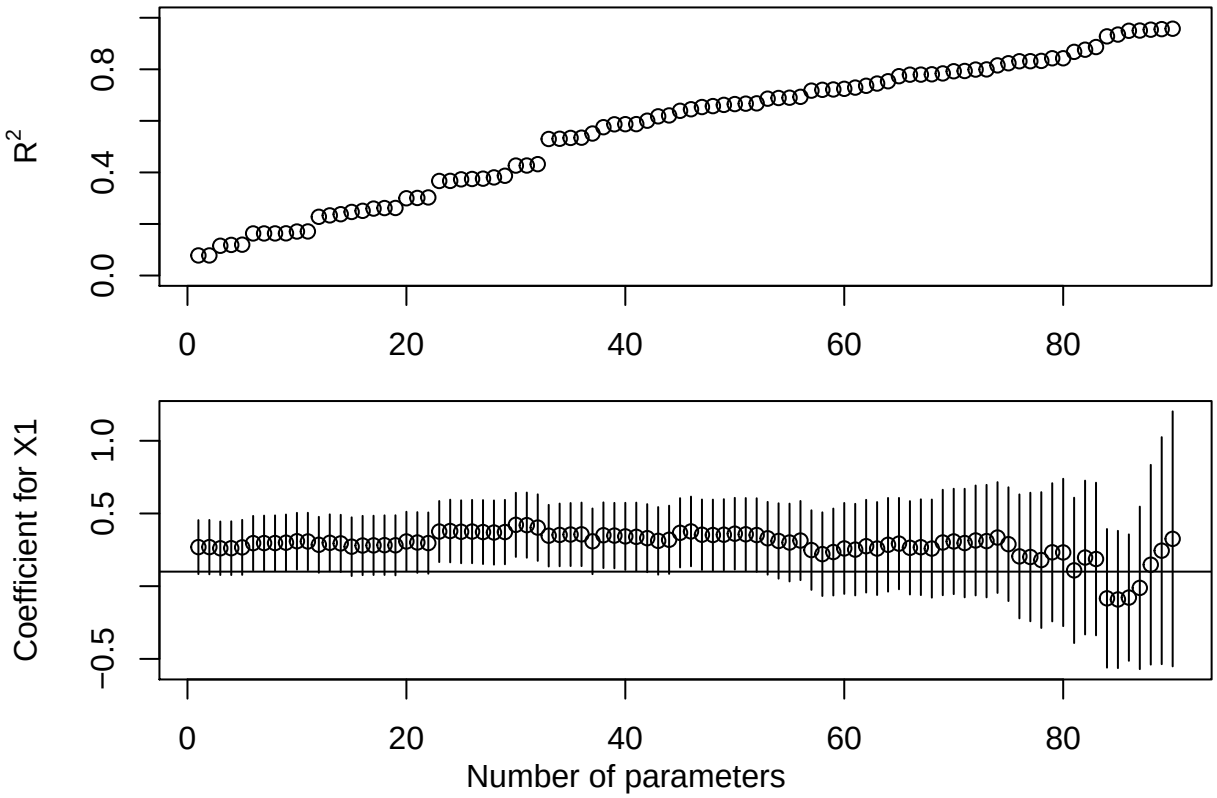
```r
source("https://www.math.ntnu.no/emner/ST2304/2021v/Week10/ModeSelectionFunctions.R")
R2 <- sapply(1:NExplanatory, GetR2, Xmat=x, Y=y)
```

Once we have these, we should plot estimate & R^2:

```r
par(mfrow=c(2,1), mar=c(2,4.1,1,1), oma=c(2,0,0,0))
plot(1:NExplanatory, R2[4,], ylab=expression(R^2), ylim=c(0,1))
plot(1:NExplanatory, R2[1,], ylim=range(R2[2:3,]),
     ylab="Coefficient for X1")
segments(2:NExplanatory, R2[2,], 2:NExplanatory, R2[3,])
abline(h=0.1)
mtext("Number of parameters", 1, outer=TRUE)
```

[1] `sapply()` loops over the first argument, and puts each element into the function. If you want to work out what's going on, try this: `sapply(1:4, function(x, n) x+n), n=2)`. See what it does, and also change `n`

Plot estimate & R^2



When we add more variables several things happen:

- R^2 increases
- parameter estimates get less precise
- interpretation can become more difficult (because there are more moving parts in the model)

So we need to balance the fit of the model to the data and the simplicity of the model. There are a few ways to do this, depending on the nature of the questions we are asking.

Two types of problem: two solutions

As noted above, there are two types of model selection problem, which we call *confirmatory* and *exploratory*.

Confirmatory model selection tests a specific hypothesis (e.g. with a t-test). The hypothesis is explicitly of interest.

Exploratory model selection is about finding a good model - something that will (for example) predict well.

Question: Which of these is exploratory & which confirmatory?

Candidate Gene Approach: does BRCA1 affect the probability of getting cancer?

GWAS: which of these 30 000 SNPs explains the probability of getting cancer?

Answers

(1) Candidate Gene Approach: does BRCA1 affect the probability of getting cancer?

This is **confirmatory**. Whe have a definite hypotehsis: that BRCA1 affects the risk of getting breast cancer, so we want to confirm if that hypothesis is correct.

GWAS: which of these 30 000 SNPs explains the probability of getting cancer?

This is **exploratory**: we are exploring the possibilities that one of the SNPs is closely linked to a region of the genome that affects breast cancer risk.

Next

Next, you should go onto the next module, about about hypothesis testing. The final module will be last, and is about exploratory model selection.