Tutorial #10: Model selection

Guy F. Sutton

Centre for Biological Control, Rhodes University

10. Model selection

In all of the previous tutorials we have been interested in *hypothesis testing*, i.e. does our response variable significantly differ or correlate with a predictor variable? However, we are not always interested in hypothesis testing, or we may not have the data to perform hypothesis testing. For example, we may be less interested in determining which variables are statistically correlated with increased cover of an invasive plant than predicting the percentage cover of that plant at different sites of differing environmental characteristics. In this case, our goal is *prediction*.

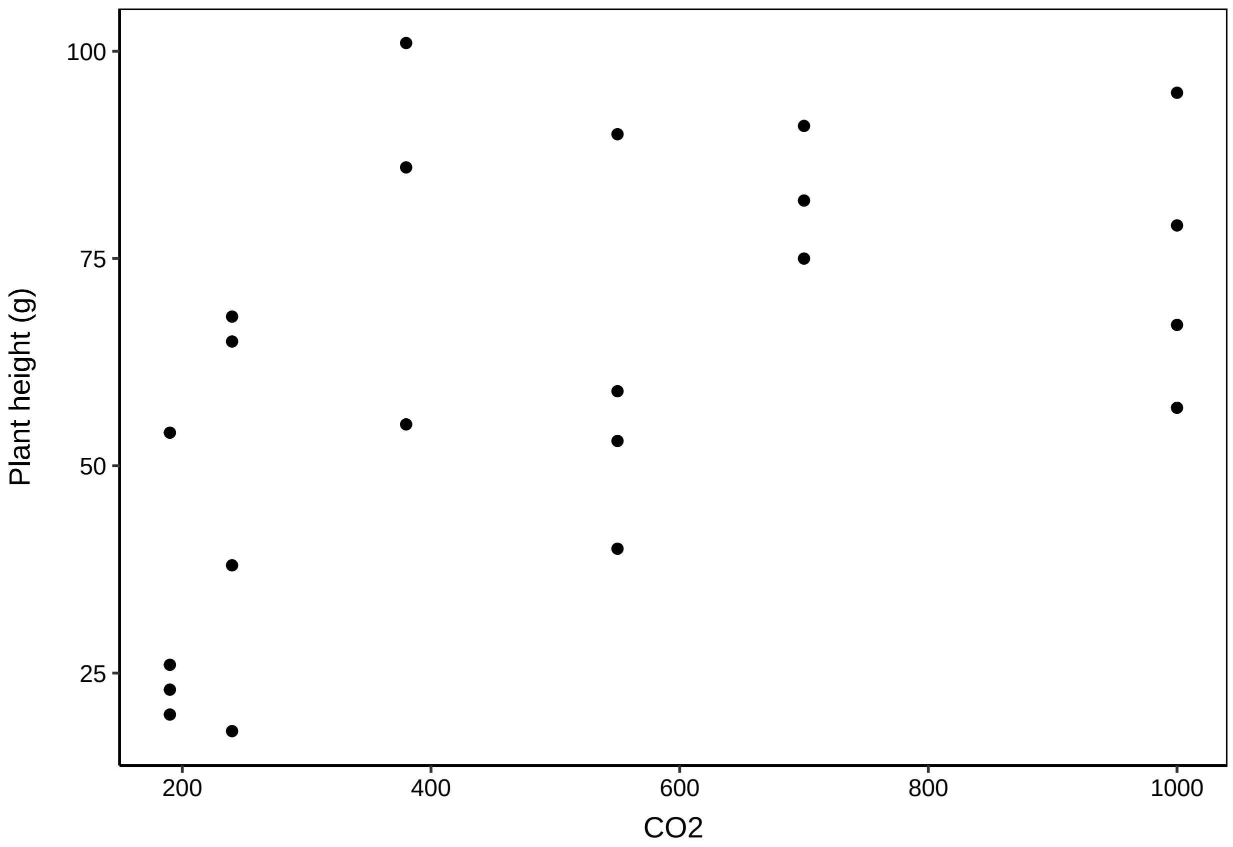
Alternatively, we could have a dataset which represents the number of seedlings at hundreds of field sites across the country where we recorded 10 climatic variables, 10 vegetation variables and 10 soil variables. It is not possible to perform hypothesis testing with so many predictor variables - the more statistical tests you perform, the more likely you are to get an incorrect result… In this case, and is often the case with observational field studies, your goal should rather be *hypothesis generation*. In other words, you can perform an analysis to identify potentially important variables that you could investigate in more detail with follow-up studies or an experimental study.

Another alternative is when we want to *compare a number of competing hypotheses* and weigh the statistical evidence for each competing hypothesis. We will work through an example of this below, although the process is very similar if we wanted to optimise predictive capacity (*prediction*) or generate testable hypotheses for future experiments (*hypothesis generation*).

10.1. An example - *Acacia* seedlings and CO2

For example, let's consider a hypothetical study where we grow *Acacia* seedlings at different CO2 levels. We come to the analysis and want to evaluate how *Acacia* will respond to increasing CO2 levels in the future. This example, code and tutorial are inspired by a seminar given by Prof. Res Altwegg at SEEC.

```r
# Plot data
data %>%
  ggplot(data = ., aes(x = CO2,
                       y = plant_height)) +
  geom_point() +
  labs(y = "Plant height (g)")
```



We read a bunch of papers and it becomes very confusing: some studies report no effect of CO2 on plant growth, some studies report plants increasingly linearly in size with increasing CO2, some studies report that plant growth initially increases with increasing CO2 but then flatlines, and lastly some studies report that plant growth decreases with increasing CO2.

Potential hypotheses We want to test these hypotheses and evaluate which one best explains our *Acacia* dataset. Let's write out our hypotheses:

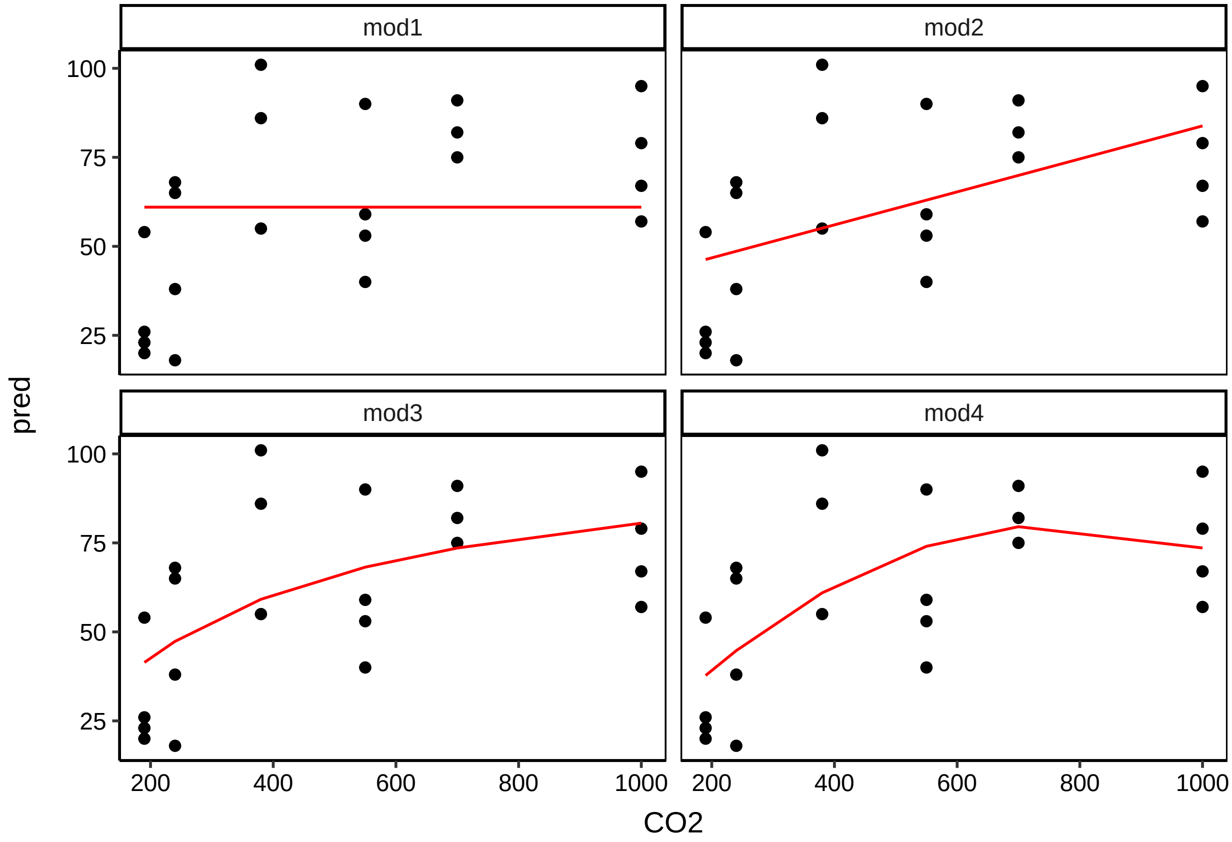
Hypothesis #1: - Plants grow to the same height, irrespective of CO2 levels. - This is equivalent to a null hypothesis (H0). - Statistical model: Intercept only model

Hypothesis #2: - If we increase CO2, plant height will increase linearly. - Statistical model: Gaussian GLM (linear regression)

Hypothesis #3: - If we increase CO2, plant height will initially increase and then growth will cease. - Statistical model: Non-linear model

Hypothesis #4: - If we increase CO2, plant height will initially increase and then growth will decrease. - Statistical model: Quadratic model

Don't worry about the different statistical models right now. Just know that we have fit 4 different statistical models to our *Acacia* data.



Which of these models/hypotheses provides the best fit to the data?

Fit the competing statistical models

Evaluating models In previous tutorials, we performed Likelihood Ratio Tests (LRT's) to calculate a p-value for predictor variables to determine whether it was statistically significant. Here, we are going to adopt a different approach where we calculate the likelihood of different models being the best fit to the data.

To do so, we calculate two different statistics: (1) Akaike's Information Criterion (AIC) and (2) Akaike weights. It is outside the scope of this tutorial to go into the derivation and mathematics of these different metrics.

- (1) *Akaike's Information Criterion (AIC)*: `AIC` balances the complexity of your models versus its predictive power. We calculate an `AIC` score for each model, and the model with the lowest `AIC` score is the best fit for our data.
- (2) *Akaike Weights (Wi)*: These represent the likelihood that each model/hypothesis is the best fit to our data. `Wi` are scaled between 0 and 1 - the closer weight is to 1, the greater the likelihood that model is the best fit to your data.

```
# Calculate AIC
aic_values <- AIC(mod1, mod2, mod3, mod4)

# Calculate delta AIC
aic_delta <- aic_values$AIC - min(aic_values$AIC)

# Calculate Akaike weights
wi <- exp(-0.5 * aic_delta) / sum(exp(-0.5 * aic_delta))

# Calculate LogLik
logLiks <- c(logLik(mod1),
             logLik(mod2),
             logLik(mod3),
             logLik(mod4))

# Define vector of model names
models <- c("mod1",
            "mod2",
            "mod3",
            "mod4")

# Combine values for a Table for a manuscript
ms_table <- data.frame(models,
                       -2*logLiks,
                       aic_values$df,
                       aic_values$AIC,
                       aic_delta,
                       wi) %>%
  dplyr::rename(Model = models,
                LogLik = 2,
                df = 3,
                AIC = 4,
                Delta_AIC = 5,
                Wi = 6)
ms_table
```

```
##    Model   LogLik df       AIC Delta_AIC          Wi
## 1  mod1 203.6960  2 207.6960  8.787230 0.006947995
## 2  mod2 196.1630  3 202.1630  3.254303 0.110489076
## 3  mod3 192.9087  3 198.9087  0.000000 0.562318452
## 4  mod4 192.0347  4 200.0347  1.125967 0.320244477
```

The model with the lowest `AIC` value (`deltaAIC` = 0) was `mod3`, so this was technically the top-performing model (i.e. the model/hypothesis with the most support). Remember, `mod3` was testing the hypothesis that if we increase CO2, *Acacia* seedling growth will initially increase and then growth will cease once we reach higher CO2 levels.

However, when we look at the Akaike Weights column (`Wi`) we can see that the likelihood of `mod3` being the best model is only 56%. There is a 32% chance (Wi = 0.32) that `mod4` was the model that best fit our data. Remember, `mod4` was testing the hypothesis that if we increase CO2, *Acacia* seedling growth will initially increase and then growth will decrease once we reach higher CO2 levels.

The general rule-of-thumb is that the `Wi` for a single model to be unequivocally the top supported model should be > 0.8 (or greater than an 80% chance of being the best model).

So, how would you go about writing up these results? Well, there are at least two options. Firstly, you present the results as they are - you found support for two different hypotheses explaining *Acacia* seedling responses to increasing CO2. Both hypotheses predict that *Acacia* seedlings will grow faster with initial CO2 increases, but once CO2 ramps up quite a bit (>700 ppm, if we look at the graph), growth may flatline (`mod4`) or it may start to decrease (`mod4`). Alternatively, you could extract predictions by averaging over the top supported models (this is called *model averaging* in the literature). Model averaging is very common, but needs to be used cautiously - does it make sense to take averages? My experience is that it very infrequently makes sense to perform model averaging, so I am purposefully not going to include it here. Send me an email if you want to perform model averaging.