6

Model Selection

6.1 What models should I fit? The candidate model set.

How to choose the model(s) you want to fit to your data is a bit of an art and it is here where your knowledge of your study system becomes critically important. Remember George Box's remark "all models are wrong but some are useful". Whether a model is useful in a particular situation depends not so much on whether it can technically be fitted to the data but on whether it helps you answer the questions that motivated you to collect these data in the first place. There is no short cut, no automatic routine, that can replace critical thinking when selecting one or more candidate models to fit to the data. It is one of the most difficult steps in data analysis, yet also one of the most rewarding because it is here where you can bring your subject knowledge to bear on the problem at hand in the most powerful way. You are confronting your theory with data.

We briefly mention a few automatic procedures that you might encounter. These methods are meant for data-mining exercises where the goal is to generate hypotheses, rather than testing them. Presenting the results of a data-mining analysis as if they were obtained through a hypothesis-driven approach is a serious error and accounts for a lot of unreproducible research in the literature. So it is really crucial to choose the correct analysis approach and be candid about the goals of your analysis when presenting your results.

6.2 Methods suitable for data mining and hypothesis generation

All subsets regression

Fit all possible models with 1, 2, ... , p explanatory variables.

The all subsets regression procedure entails fitting all possible models: those with 1, 2, ..., p explanatory variables (there are

$2^{p-1} - 1$ possibilities). From this set of models we choose the one that fits best according to some criterion (adjusted R^2, AIC, Mean square for error (MSE), or Mallows C_p statistic).

The problem with this approach is that one quickly ends up fitting a large number of models to a limited data set. For example with 10 explanatory variables, you would fit 511 models. We have seen cases where researchers fitted hundreds of models to less than 20 data points. This approach is guaranteed to lead to overfitting and the problem known as Freedman's paradox where predictor variables with no explanatory power appear artificially important. At best, this approach can generate hypotheses about relationships that need to be tested on independent data sets.

Stepwise Regression Procedures

There are a number of iterative model fitting procedures that you will encounter in the literature. In one variant, one starts with the full model (containing all the variables available) and then eliminates variables in order of their significance (the one with the highest p-value first) until only variables with small p-values remain in the model. This is called backward stepwise regression. Forward stepwise regression starts from the null model and adds variables one at a time, always the one with the smallest P value first. Finally, there is a procedure that combines forward and backward stepwise regression.

The problem with stepwise regression is again overfitting, i.e. you are almost guaranteed to find some spurious results: variables with no predictive power appear statistically important.

These automated model selection procedures are quick and easy to use. For example, there is an R function, 'dredge' in package 'MuMIn' that does it automatically for you. However, these methods should not be used for testing scientific hypotheses because they lead to overfitting and spuriously significant results. The main purpose of these procedures is for data-mining analyses, i.e. if you really have no a priori hypotheses about the processes that could have generated your data. This seems to us to be a rare situation in research and the results are weak: you need to treat the patterns you find as hypotheses that should be tested on independent data before you can claim that they are real. We see two uses for these automated model selection routines: the first is for generating hypotheses as mentioned above; the second is after you have conducted a rigorous model selection analysis and you want to explore further patterns in the data that no-one expected. In either case, you clearly need to declare these kinds of analyses as exploratory data analysis and interpret the results accordingly.

6.3 *The problem with multiple testing*

If we test a lot of null hypotheses that are in fact true (e.g. explanatory variables that have no true relationship with the response variable), a few of them will look unlikely (small p-value) just because the data happened to be unusual. We expect to find 5% of our p-values to be <0.05, 1% to be <0.01, and so on, just by chance. So if we test 100 variables and find that 5 of them lead to a $P < 0.05$, we probably haven't found any true relationships at all. This problem is most obvious in experiments when a number of different treatments are compared, in which case the number of tests can quickly approach vast numbers, e.g. if there are 5 different drugs and I would want to make all pairwise comparisons there are $\binom{5}{2} = 10$ hypothesis tests or confidence intervals. We will dwell a bit more on this problem in the Experimental Design module of this course.

If a multiple regression with 10 explanatory variables is approached without thought, the problem is even worse: there are all possible models with 1 explanatory variable, all with 2 explanatory variables, etc., adding up to 1023 possible models (excluding models with interactions)! Without careful a-priori thinking, we are often tempted to include 10 explanatory variables in an analysis.

The problem of multiple testing can also occur in a single model with many explanatory variables. For example, we could fit a regression model with 20 explanatory variables. The regression output gives a t-test and related p-value for each of these. How does one ensure that one does not pick up spurious relationships (by chance), but only real relationships?

The solution is discussed below: before starting a project that will involve statistical analyses the first step should always be: Think hard: What do you want to know? What do you think is important? Based on this, only consider the variables that you really think might be important and construct a small set of candidate models.

6.4 *Model selection for hypothesis-driven research*

Carefully constructed candidate model set

We advocate putting together a small set of candidate models to fit to your data set. The set should consist of all models that you are interested in, based on your understanding of the system but it should be small relative to the number of data points. The idea is that each model represents an alternative hypothesis about the processes that generated the data and you should be able to justify the inclusion of each model. This is probably the most important

step in scientific data analysis and one of the more time consuming parts. You should ideally decide on the candidate model set before you start collecting data but definitely before you start looking at the data.

The effort you spend choosing a good candidate model set is well invested. Most importantly, the results of your model selection analysis will be straightforward to interpret: you will have a measure of the degree to which the data support each of your models and the hypotheses that motivated you to fit each model.

Let's look at an example, motivated by a study on the effect of CO_2 level on growth of sweet thorn (*Acacia karroo*), a savanna tree. Sweet thorn is one of the species involved in bush encroachment that currently affects large parts of southern Africa where woody vegetation is becoming denser and in areas replaces open vegetation. One of the causes for bush encroachment could be the globally increasing CO_2 concentration that might allow these trees to grow more quickly and outcompete other vegetation types (e.g. grasses). How would you test this hypothesis?

Kgope et al. (2009)[1] grew sweet thorn under different levels of CO_2 (at 180, 280, 370, 550, 700 and 1000 ppm) and measured the stem length after six months. They had four replicates of each treatment. Imagine you are planning to conduct this experiment. The first step would be to come up with a set of hypotheses on how stem length and CO_2 level are related, and translate them into statistical models to be fitted to the data. After some thinking, we came up with the following four hypotheses:

[1] B. S. Kgope, W. J. Bond, and G. F. Midgley. Growth responses of African savanna trees implicate atmospheric CO_2 as a driver of past and current changes in savanna tree cover. *Austral Ecology*, 35:451–463, 2009

1. CO_2 has no effect on plant height after 6 months; perhaps another factor is limiting growth.
2. Plant height increases linearly with increasing CO_2 concentration.
3. The relationship between height and CO_2 is a saturation curve: as CO_2 becomes higher, its effect diminishes.
4. There is an optimal level of CO_2 that leads to maximum height. Lower and higher levels lead to slower growth.

Figure 6.1 shows a sketch of the hypothesized relationships.

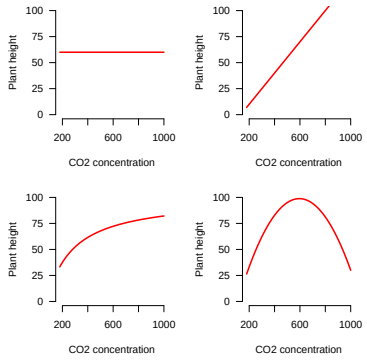

Figure 6.1: Possible relationships between stem length of *Acacia karroo* and CO_2 concentration.

We then formulate a model that can structurally represent each of these hypotheses. The models we are going to fit are the following, where 'H_i' stands for the measured height of plant i, 'CO2' is the CO_2 concentration, the β's are parameters to be estimated and the ϵ_i are normally distributed errors following $N(0, \sigma^2)$:

1. An intercept-only model, i.e. the slope is constrained to be zero: $H_i = \beta_0 + \epsilon_i$. This model has just two parameters: the intercept and residual standard error, σ.

2. A linear relationship can be described using a linear regression model: $H_i = \beta_0 + \beta_1 \times CO2_i + \epsilon_i$. This model has three parameters: the intercept, slope and residual standard error, σ.

3. We use a non-linear regression model to describe a saturating relationship: $H_i = \beta_0 \times CO2_i/(1 + \beta_0 \times \beta_1 \times CO2_i) + \epsilon_i$. This model has three parameters: β_0 and β_1 together determine the shape of the curve and the residual standard error, σ.

4. We add a quadratic term to the linear regression model to allow the fitted relationship to assume an optimum: $H_i = \beta_0 + \beta_1 \times CO2_i + \beta_2 \times CO2^2 + \epsilon_i$. This model has four parameters: the intercept, coefficient for the linear term, coefficient for the quadratic term and residual standard error, σ.

In R, we fit these models using the following code:

```
m1<-lm(Total.stem.length ~ 1)
m2<-lm(Total.stem.length ~ CO2)
m3<-nls(Total.stem.length ~ (beta0 * CO2) / (1 + beta0 * beta1 * CO2),
        start = list(a = 1, b = 1 / 100))
m4<-lm(Total.stem.length ~ CO2 + I(CO2^2))
```

Models 1, 2 and 4 are linear models and we can fit them using function `lm()`. Model 3 is a nonlinear model and we fitted it using function `nls()`. While the parameters (coefficients) are implicit in the notation for linear models, with `nls()` we have to write out the whole model, including the two structural parameters, which we called `beta0` and `beta1`. We also have to supply reasonable starting values for these parameters because `nls()` uses an iterative fitting algorithm.

The fitted regression lines are shown in Figure 6.2. Which of our four hypotheses is best supported by the data? Since each model corresponds to one hypothesis, another way to answer this question is to ask which of our four models is best supported by the data. How to do this is the subject of the next sections.

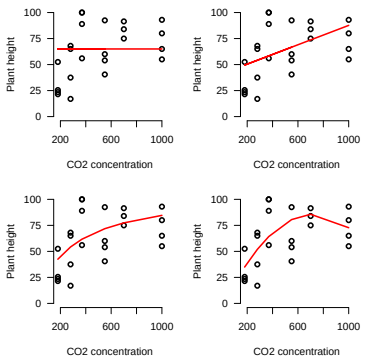

Figure 6.2: Observed data and fitted regression lines for stem length vs CO_2 level of *Acacia karroo* experiment.

6.5 *Finding the best model: model comparison and selection*

When we have several models all attempting to describe our data or to make predictions, how do we choose among them? Among explanations (think models) that fit the facts (think data) similarly well (i.e. they explain the same amount of variability in the response), a simpler explanation is better than a more complex one. This principle is known as parsimony or Occam's razor. However, if a more complex model explains the observed data much better, this more complex model should be preferred.

To choose a parsimonious model we trade off goodness-of-fit and number of parameters used. We now discuss five methods to choose between models. Of these we prefer Akaike's Information Criterion (AIC), as this is the most flexible and is easily extended to other types of models, such as generalized linear models (Module 3). We largely follow the approach in Burnham and Anderson (2002)[2].

[2] Kenneth P Burnham and David R Anderson. *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*. Springer, 2002

1. The **adjusted** R^2 (Section 3.5) is one measure that lets us compare models. For the linear models in the CO_2 example, the R^2 and adjusted R^2 values are:

```
# m1 (trivial...)
summary(m1)$r.squared; summary(m1)$adj.r.squared

# m2
summary(m2)$r.squared; summary(m2)$adj.r.squared

# m4
summary(m4)$r.squared; summary(m4)$adj.r.squared
```

   For non-linear models, R^2 has a slightly different interpreation(!), which is why `nls()` doesn't automatically calculate it and we need to do a bit more work (there is no comparable adjusted R^2 for nonlinear models):

```
# m3:
(RSS <- sum(residuals(m3)^2))  # Residual sum of squares
# Total sum of squares:
(TSS <- sum((Total.stem.length - mean(Total.stem.length))^2))
1 - (RSS/TSS)  # R-squared
```

   As expected, the most complex model, m4, has the highest value for R^2 (Table 6.1). This does not necessarily mean that it is the best model for prediction. It just means that this model explains the highest proportion of variance in plant height, among the models we considered. However, R^2 is a good tool to help us judge how well the model fits the data, i.e. to judge goodness-of-fit. Both models 3 and 4 explain just over one third of the variance (34% and 41%, respectively). Whether this is adequate depends on what your purpose is with these models. However, these figures are fairly typical for ecological data sets.

Table 6.1: Comparison of four CO_2 models by R^2, adjusted R^2 and residual mean square (MSE).

Model	R^2	adjusted R^2	MSE
1	0	0	768.83
2	0.22	0.19	622.89
3	0.34		530.68
4	0.41	0.35	498.63

2. The **residual mean square** (or mean squares for error, MSE) estimates the residual variance (unexplained variation), σ^2_ε. This should decrease as more important variables enter into the regression equation. MSE will tend to stabilise as the number of variables included in the equation becomes large. The model that

minimises the *MSE* fits the data most closely. The MSE is directly related to the adjusted R^2. The R output obtained through the `summary()` function shows the residual standard error, which is the square root of MSE.

```
summary(m1)$sigma^2
summary(m2)$sigma^2
summary(m3)$sigma^2
summary(m4)$sigma^2
```

3. **Mallow's C_p statistic** is an estimate of the *prediction error*, which is a combination of bias and precision. A good model should predict well. For a well fitting model C_p should be close to k, the number of β parameters the model has estimated (including the intercept term). For linear models, C_p is equivalent to Akaike's Information Criterion (AIC), which we will discuss in detail below.

4. **Analysis of Variance / Deviance** can be used to compare nested models. A model A, say, is *nested* in model B if all terms in model A also appear in model B, but not all terms of model B need to be in model A (model A is the simpler model, model B has some extra terms). In our example, m1 is nested in m2 (forcing the slope, β_1 in m2 to be equal to zero yields m1) and m1 and m2 are nested in m4. However, m3 is not nested in any of the other models, nor is any of the other models nested in m3.

   Analysis of deviance examines the *change in the amount of variance explained*. If this is large relative to the number of extra parameters estimated, then model B (the more complex model) is better, else, the simpler model is preferred. We can compare the Regression Sum of Squares (RSS) (amount of variation explained) of the two models. Whether the difference (or change) in RSS is significant can be tested with an F-test: does the extra term in the model help to explain sufficiently more of the variation in Y than the simpler model?

   Let's compare m1 and m2 to test whether CO_2 affects plant height:

```
> anova(m1,m2)
Analysis of Variance Table

Model 1: Total.stem.length ~ 1
Model 2: Total.stem.length ~ CO2
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     23 17683
2     22 13704  1    3979.5 6.3889 0.01917
```

   For one extra parameter (degree of freedom) the residual sum of squares decreases by 3979.5, comparing that to the MSE gives an

F-statistic of $6.4 \sim F_{1,22}$. The small p-value ($P = 0.019$) indicates that this change is unlikely under the null hypothesis (extra parameters are all 0), and therefore there is evidence that the extra parameter improves the model (improves the amount of variation explained).

The R output here may be slightly confusing because RSS here stands for *residual* and not regression sum of squares. The residual SS will always decrease when more parameters are added to a model, but this decrease may not be worth the 'cost' of the extra parameters.

```
> anova(m1,m2,m4)
Analysis of Variance Table

Model 1: Total.stem.length ~ 1
Model 2: Total.stem.length ~ CO2
Model 3: Total.stem.length ~ CO2 + I(CO2^2)
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     23 17683
2     22 13704  1    3979.5 7.9809 0.01014
3     21 10471  1    3232.2 6.4822 0.01881
```

Model comparison using ANOVA only works for nested models, so m3 can't be included in this comparison. Following the same logic as when we compared just m1 and m2, we see that m4 is also an improvement over m2. From these pair-wise comparisons, we conclude that m4 is the best model of the ones we considered. (The p-value and F-statistic for m2 have changed slightly. This this is because the MSE used is always that of the largest model, which is now m4, not m2 as before).

You can see that model-selection by Analysis of Variance is fairly cumbersome if you compare more than just two models. Also the fact that you can only compare nested models restricts your creativity as a modeller. However, this method is sometimes useful if you want to compare two specific models in your model set.

5. **Information Criteria**

   The basic problem that a model selection tool needs to solve is how to best balance the trade-off between overfitting and underfitting. Overfitting happens when a model is too complex for the data set in question. The model fits to noise in the data, rather than describing the underlying pattern. Overfitting also leads to large standard errors because each parameter estimate is based on few data points. A model that overfits is poor at predicting new data because of the large uncertainty in the parameter estimates.

   Underfitting, on the other hand, happens when a model is too simple – think too rigid – to describe the structure in the data

adequately. Underfitting leads to bias. Because we only have to estimate a few parameters, we can get very precise estimates. This is not a good thing in this case because we would be too confident in our parameter estimates. A model that underfits is also poor at predicting new data. We may get precise predictions but they will often be precisely wrong.

So there is a trade-off between bias and variance related to model complexity. A simple model tends to underfit and make biased predictions whereas a complex model tends to overfit and make variable predictions. We need a model selection tool that balances this trade-off and identifies the model that best describes the structure in our data. Akaike's Information Criterion (AIC) is such a model selection tool.

AIC can be used to compare nested or non-nested models (see Appendix 1 for some background on likelihood and the AIC). It is calculated as:

$$AIC = -2\log(\mathcal{L}(\hat{\theta}|\text{Data})) + 2K$$

Here $\log(\mathcal{L}(\hat{\theta}|\text{Data}))$ is the maximized log-likelihood for a particular model, and K is the number of estimated parameters. The $-2\log(\mathcal{L}(\hat{\theta}|\text{Data}))$ is a measure of distance from a best possible model. You can think of it as a measure of how closely the model fits to the data. The larger $-2\log(\mathcal{L}(\hat{\theta}|\text{Data}))$ (minus 2 times log-likelihood) is, the worse the model; the smaller $-2\log(\mathcal{L}(\hat{\theta}|\text{Data}))$ is, the better the model. You can think of the $2K$ as a penalty for model complexity, because the more parameters we use in a model (i.e. the more complex the model is) the better the model will be able to explain the data, even if the parameters are not related to the response.

The model with the *smallest AIC value* in the set is therefore the best model. We next calculate the AIC value for the four models that we fitted to the sweet thorn data:

```
> aics <- AIC(m1, m2, m3, m4)
> aics
   df      AIC
m1  2 230.56
m2  3 226.45
m3  3 222.60
m4  4 221.99
```

Model m4 has the smallest AIC value and is therefore the best model in our set. However, the AIC value for model m3 is not much higher. How much better is model m4 than model m3?

The absolute value of the AIC is not informative, it depends on the data set. Only the difference or change in AIC when comparing models is of interest. We therefore next calculate ΔAIC, the difference in AIC between each model and the best.

```
delta.aics <- aics$AIC - min(aics$AIC)

cbind(model=c("m1","m2","m3","m4"), delta.aics)

     model delta.aics
[1,] "m1"  8.58
[2,] "m2"  4.46
[3,] "m3"  0.61
[4,] "m4"  0.00
```

Looking at the ΔAIC makes it clear that models 3 and 4 are close competitors for being the best model in the set. At this stage, it would be useful to know how likely each model is, given the data. As it turns out, the likelihood of a model g_i is proportional to a quantity that we can easily derive from the ΔAIC values:

The delta symbol Δ is often used to denote *change*.

$$\mathcal{L}(g_i|x) \propto exp(-\frac{1}{2}\Delta_i)$$

These likelihoods represent the relative strength of evidence for each model. To make them more easily interpretable, we scale these values to sum to 1.

$$w_i = \frac{exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R} exp(-\frac{1}{2}\Delta_r)}$$

These scaled values are called **Akaike weights**, w_i, and are interpreted as strength of evidence for a particular model to be the best, relative to the other models in the set. Even though the equation above may look a bit daunting, the calculation is actually quite simple and can be carried out with one line of R code:

```
wi <- exp(-0.5*delta.aics)/sum(exp(-0.5*delta.aics))

cbind(model=c("m1", "m2","m3","m4"),wi)

     model wi
[1,] "m1"  0.01
[2,] "m2"  0.06
[3,] "m3"  0.40
[4,] "m4"  0.54
```

Up to some rounding error, these Akaike weights sum up to 1 across the models in the set.

Model m4 has 54% of the total weight and model m3 has 40%. We can go one step further and calculate **evidence ratios**, i.e. the odds that one model is in fact the best, compared to another model. Comparing model m4 to model m3, the evidence ratio is $\frac{w_4}{w_3} = \frac{0.54}{0.40} = 1.35$, i.e. model 4 is only 1.35 times more likely to

Some situations where AIC does not work:

- All models must be based on exactly the same data points. One situation that can easily make AIC-based model selection invalid is if there are missing values for some of your explanatory variables. Let's assume you have n observations Y, which you want to regress against covariates A, B and C and one of the observations is missing information for covariate C. By default, R (and most other statistical programs) omits this observation when fitting a model that involves covariate C. So for example the model $Y \sim A + B + C$ will be fitted to $n - 1$ data points whereas the model $Y \sim A + B$ will be fitted to n data points. These two models cannot be compared using AIC because they are not based on the same data set. R removes observations with missing values without complaining so you may not notice that anything went wrong. If you have missing values in your data set, it is safer to remove all rows with missing data before you start the analysis.
- By changing the response variable, the likelihoods become incomparable, and then AIC-based model selection is invalid. For example, you cannot compare model $Y \sim A$ to model $\log(Y) \sim A$.

be the best model than model 3. Comparing model 4 to model 2 $\frac{w_4}{w_2} = \frac{0.54}{0.06} = 9$, we see that model 4 is 9 times more likely than model 2.

If these models were race horses, you probably would not put all your money on one of them. You would be quite confident that either m4 or m3 wins but it is not clear which one of them is the better one (they have only competed in one race so far). Likewise, the interpretation of the results of this analysis is that models m4 and m3 are close competitors and the data do not allow us to clearly distinguish between the two hypotheses they represent. This is not a deficiency of AIC but rather an indication that the data are ambivalent about these two structures.

In a report, we would say that there is clear evidence that CO_2 concentration affected plant height and that the effect levelled off at high CO_2 concentrations. However, the data do not clearly show whether very high CO_2 concentrations depress plant height or not. This kind of narrative needs to be supported by a model selection table (Table 6.2) that allows readers to verify your conclusions. The table should include at least the model names (or description), $-2\log(\mathcal{L}(\hat{\theta}|\text{Data}))$, the number of parameters (often labelled 'K'), Δ AIC, and the Akaike weights. Most model selection tables also report the absolute value of AIC, even though this value is not of direct interest and could easily be calculated from the other information in the table.

	-2 × loglik	K	AIC	Δ AIC	w
m1	226.56	2	230.56	8.58	0.01
m2	220.44	3	226.45	4.46	0.06
m3	216.60	3	222.60	0.61	0.40
m4	213.99	4	221.99	0.00	0.54

Table 6.2: Model selection table for *Acacia karroo* data.