Model Selection

- Models are meant to be a representation of the real world, they are:
  - Imperfect – always some uncertainty
  - Simplified – goal is to represent as much of the real world as possible in a way that is understandable
  - Tradeoff between model fit and model parsimony

**Model Selection**

• Given the observed data, we can quantify model fit in a number of ways
  - R^2
  - RMSE
  - Maximum Likelihood – the most likely set of parameters and model given the data

10 heads from 30 tosses of a coin, 10/30 = 0.3 \( \rightarrow \) MLE

\[
G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{8\pi G}{C^A} T_{\mu\nu}
\]
Model Selection

Model complexity

- Underfit
- Overfit

Parsimony

- Information

"When multiple competing hypotheses are equal in other respects, the principle recommends selecting the hypothesis that introduces the fewest assumptions and postulates the fewest entities."

Overfit
- Include unnecessary variables
- Lack of parsimony
- Difficult interpretation

Underfit
- Lack important variables
- Minimal explanatory power
- Inaccurate parameter estimates

Goal is to strike a balance between these two.

Parsimony – Fit in CCA

- We used similar logic earlier in CCA
- Stepwise regression and VIF score examination were aimed at eliminating redundancy (avoid overfit)
- Selecting the best variables and maximizing % variance explained (avoid underfit)

Model Selection and Parsimony in Ecology

- Much of ecology is observational
- Hypotheses often of the form: "which of these observed variables best explains the observed ecological process or pattern"

- Traditional approach
  - $H_0$: response variable $y$ differs significantly with respect to variables a, b and c
  - Test the significance of $H_0$

- Information-theoretic approach
  - What set of observed variables form the most parsimonious and accurate model predicting the response variable
  - No significance test
Akaike Information Criterion – Parsimony vs. Fit

- Mathematical estimate of the balance between parsimony and fit of a model.

\[ AIC = -2 \log(L(model|data)) + 2k \]

- Minimizing AIC = finding the set of explanatory variables that explain the most about the response variable(s).
- Addition of parameters is penalized

**Uses**
- Compare models (each model represents a competing hypotheses)
- Compare levels of treatments

**AIC in R**
- Function `AIC()` will calculate AIC for any model
- Simple application
  - Response variable `A`
  - Factors `B`, `C` and `D`
  - Compare AIC values for models
    - `A ~ B*C*D` (global)
    - `A ~ B*C`
    - `A ~ B*D`
    - `A ~ B`
    - `A ~ 1` (null)

- The absolute size of AIC values relative to other competing models is what matters.
- AIC values not comparable across studies

**Akaike Information Criterion**


- There are modifications to this (weights for variables, sample size adjustments etc.), but the general idea is the same.
  - AIC_c
  - BIC

- Function `AIC()` will calculate AIC for any model
  - `model`
  - `data`

- Log likelihood of model given the data
- Model complexity

- Compare models (each model represents a competing hypotheses)
- Compare levels of treatments

- AIC values not comparable across studies
AIC in Ecology

• Ask good questions
• Conduct well designed studies
• Develop a set of plausible models to explain data
• Calculate AIC values for this set of models
• Rank models by AIC value (lowest to highest)
• Compute the difference in AIC values between the first and each subsequent model
  - AIC difference $\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}}$
  - $\Delta_i$ values: 0-2 = good support, 4-7 = less support, >10 none
• Remember, AIC values within a set of competing models are only comparable among themselves

Model Comparison, example dataset

• NMDS of community data
• What environmental data best predicts axis 1 scores?
• Global model:

```
Call: lm(formula = metanmds$points[, 1] ~ ., data = envdata)

Residuals:
Min 1Q Median 3Q Max
-0.37338 -0.11365 -0.01083 0.14677 0.25822

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -3.135265   0.427212 -7.339 3.68e-06 ***
spatialb                  0.178619   0.243453  0.734   0.4752
spatialc                  0.148344   0.444862  0.333   0.7437
env1                      -0.300768   0.261051 -1.152   0.2686
env2                      -0.544517   2.419253 -0.225   0.8252
env3                       0.292868   0.165836  1.766   0.0992 .
env4                       0.003002   0.227521  0.013   0.9897
env5                      -0.046674   0.222875 -0.209   0.8371
env6                      -0.192800   0.222696 -0.866   0.4012
env7                       6.516912   0.722254  9.023 3.29e-07 ***
env8                       0.018205   0.233328  0.078   0.9389
env9                       0.169447   0.181752  0.932   0.3670
env10                     0.488443   2.456476  0.199   0.8452
env11                     -0.246350   0.198231 -1.243   0.2344
env12                     -0.211397   0.200459 -1.055   0.3095

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.24 on 14 degrees of freedom
Multiple R-squared: 0.9906, Adjusted R-squared: 0.9813
F-statistic: 105.7 on 14 and 14 DF,  p-value: 1.037e-11
```

• Competing models (each should represent a hypothesis):
  - Null – no variables
  - Global – all variables
  - Spatial variable
  - Environmental variable 7
  - Spatial and environmental variable 7

Function AICtab in package bbmle
Classification Trees and Regression Trees

- **Classification Trees**
  - Descriptive technique
  - Multiple samples are categorized (by treatment, location, presences of species etc) into A-F.
  - Tree is built (recursive partitioning)
  - Eg. goal is to describe differences in samples A-F in relation to predictor variables (P1-P5)

- **Regression Tree**
  - Continuous response variable. Samples not in discrete groups.
  - At each node, analysis seeks cutoff values that group samples based on similarity in response variable.
  - Accuracy measured as homogeneity within tips.
  - Reports number of samples classified at the end of each branch.
  - Examples
    - Community productivity measured at sites. Want to describe what species (P1-P5) best predict productivity levels.

Classical Algorithm, at each iteration:
- Continuous variables transformed to ranks (non-parametric)
- First node is formed by:
  - Partitioning the data by levels of each variable
  - Choose the one that partitions to produce the most homogeneous groups.
- Within each of the two existing clusters, repeat the above process
- Iterative procedure, repeat until additional partitioning does not reduce within group heterogeneity
- One variable can appear multiple times
Classification Tree and Regression Trees

- **Function** `rpart` (**rpart** package)
  ```r
  rpart(metanmds$points[,1]~spatial+env1+env2+env3+env4+env5+env6+env7, data = envdata)
  ```

- **Formula** specifies the response variable (continuous for regression trees, a factor for classification trees) and the independent variables.

- **Options**
  - **Size**: specify number of tree nodes (if not specified, tries to guess at stopping point to avoid overfitting)
  - Numerous options for controlling when the iterations stop and how clusters are formed. See `?rpart` for details.

Boosted Regression Trees

- **Boosting** (aka machine learning) – a series of simple but poor models (recall parsimony-fit tradeoff) combined are better than a single good model.

- Boosted algorithms common in informatics (search engines, speech recognition etc.)

- Boosted regression trees (package gbm)
  - Series of simpler (more parsimony, less fit) trees combined

- **Iterative process as before…**
  - Build a regression tree as before
  - Calculate residual variation, this represents samples in the analysis that are not classified in the first tree
  - Build a regression tree classifying residuals (thus directly focusing on aspects of the dataset not addressed in first tree)
  - Combine trees
  - Repeat last three steps until a stopping point is reached.

- Often run with cross-validation (randomly select a subset of data for each iteration)
**Boosted Regression Trees**

- Control over the iterative process:
  - Tree complexity – how many nodes at each iterative step
  - Learning rate – contribution of each tree to the overall model.
  - Number of trees – how many trees can be combined in the overall model.
- Function `gbm` in package `gbm`

```
gbm(pascagoula$Cond ~ Flow.upper + Flow.lower + Temp. + Seechi + DO + Depth + pH + Month + River + Site + Habitat, data=pascagoula, distribution="gaussian", n.trees=5000)
```

**Reading**

- Text: section 4.11
- Papers
- Sample script
  - AIC model comparison (example dataset)
  - Regression and classification trees (example dataset and Pascagoula River habitat data)
  - Boosted regression tree (Pascagoula river habitat)

**Homework**

- Ecological traits as predictors of stream/reservoir abundance
  -Goldstein and Meador (2004) – 32 traits for 429 N. American species

| Species | log_resratio | Herb_det | plank | omni | invert | carniv | bedrock | boulder | NOTSHU | SEMATR | FUNZEB | NOTGRE | NOTRUB | FUNOLI | ETHZON | NOTSTR | HYBPLA | HYPNIG | LYTUMB | NOTPOT | PIMPRO | LEPMAR | ETHRAD |
|---------|--------------|----------|-------|------|--------|--------|---------|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| NOTSHU  | 0.000221     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| SEMATR  | 0.000291     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| FUNZEB  | 0.000525     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| NOTGRE  | 0.000601     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| NOTRUB  | 0.000651     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| FUNOLI  | 0.000801     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| ETHZON  | 0.000906     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| NOTSTR  | 0.000799     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| HYPNIG  | 0.000623     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| LYTUMB  | 0.000842     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| NOTPOT  | 0.001976     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| PIMPRO  | 0.001851     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| LEPMAR  | 0.002535     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |
| ETHRAD  | 0.003688     | 0        | 0     | 0    | 0      | 0      | 0       | 0       | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |

| 104 | Log ratio of mean reservoir to stream abundance. Higher=species more abundant in reservoirs. | 30 years of community data from 28 reservoirs | 30 years of community data from 208 streams and rivers | Goldstein and Meador (2004) – 32 traits for 429 N. American species | Log ratio of mean reservoir to stream abundance. Higher=species more abundant in reservoirs. | Homework | Homework

- Ecological traits as predictors of stream/reservoir abundance
- Final matrix to analyze
  - Species by ecological traits matrix
  - Response variable – log(resratio) = the ratio of mean reservoir abundance to mean stream abundance
- What traits are the best predictors of reservoir abundance?
- Use AIC to compare at least 5 models (including one null and one global)
- Regression tree (boosted or non-boosted)
I added a file that has full species names. You can use this however you want.

<table>
<thead>
<tr>
<th>Code</th>
<th>Full Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOTSHU</td>
<td>NOTROPIS SHUMARDI</td>
</tr>
<tr>
<td>SEKMAR</td>
<td>SEMOTILUS ATRIMACULATUS</td>
</tr>
<tr>
<td>FUNZEB</td>
<td>FUNDULUS ZEBRINUS</td>
</tr>
<tr>
<td>NOTGRE</td>
<td>NOTROPIS GREESEI</td>
</tr>
<tr>
<td>NOTRUB</td>
<td>NOTROPIS RUBELLUS</td>
</tr>
<tr>
<td>FUNOL</td>
<td>FUNDULUS OLIVACEUS</td>
</tr>
<tr>
<td>ETHZON</td>
<td>ETHEOSTOMA ZONALE</td>
</tr>
<tr>
<td>NOTSTR</td>
<td>NOTROPIS STRAMINEUS</td>
</tr>
<tr>
<td>HYBPLA</td>
<td>HYBOGNATHUS PLACITUS</td>
</tr>
<tr>
<td>HYPNIG</td>
<td>HYPENTELIUM NIGRIGANS</td>
</tr>
<tr>
<td>LYTUMB</td>
<td>LYTIRHUS UMBRATILIS</td>
</tr>
<tr>
<td>NOTPOT</td>
<td>NOTROPIS POTTERI</td>
</tr>
<tr>
<td>PIPIFRO</td>
<td>PIMENPHILUS PROMNEAS</td>
</tr>
<tr>
<td>LEMMAR</td>
<td>LEPOMIS MARGARITUS</td>
</tr>
<tr>
<td>ETNERM</td>
<td>ETHEOSTOMA RANGIOLM</td>
</tr>
<tr>
<td>PERCI</td>
<td>PERCINA SCIRRA</td>
</tr>
<tr>
<td>CARAUR</td>
<td>CARASSIUS AURATUS</td>
</tr>
<tr>
<td>CYCLIT</td>
<td>CYPRINELLA LUTRENSIS</td>
</tr>
<tr>
<td>MOCHY</td>
<td>MOCHILA ERYTHRURUS</td>
</tr>
<tr>
<td>LUCAR</td>
<td>LUXILUS CARDINALIS</td>
</tr>
<tr>
<td>NOTBOO</td>
<td>NOTROPIS BOOPS</td>
</tr>
<tr>
<td>PHEMIR</td>
<td>PHENACOBUS MIRABILIS</td>
</tr>
</tbody>
</table>

Ecological/Life History traits for each species by category

<table>
<thead>
<tr>
<th>Diet</th>
<th>Habitat</th>
<th>Original Habitat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herb_det</td>
<td>Riffle</td>
<td>sn_ck</td>
</tr>
<tr>
<td>Plank</td>
<td>Pool</td>
<td>sn_riv</td>
</tr>
<tr>
<td>Omni</td>
<td>Run</td>
<td>med_riv</td>
</tr>
<tr>
<td>Invert</td>
<td>Backcast</td>
<td>lg_riv</td>
</tr>
<tr>
<td>carniv</td>
<td>var_hab</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Substrate</th>
<th>Locomotion</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bedrock</td>
<td>Cruise</td>
<td></td>
</tr>
<tr>
<td>Boulder</td>
<td>Accel</td>
<td></td>
</tr>
<tr>
<td>Cobble</td>
<td>Hugger</td>
<td></td>
</tr>
<tr>
<td>Gravel</td>
<td>Creep</td>
<td></td>
</tr>
<tr>
<td>Sand</td>
<td>Maneu</td>
<td></td>
</tr>
<tr>
<td>Mud</td>
<td>Reproduction</td>
<td></td>
</tr>
<tr>
<td>Veg</td>
<td>Broad</td>
<td>Nest</td>
</tr>
<tr>
<td>var_sub</td>
<td>Bearer</td>
<td>Bearer</td>
</tr>
</tbody>
</table>
