# Model selection

Day 3: Model uncertainty and model averaging

# Model uncertainty

- Usual approach: treat the selected model as the «best» model and ignore the selection process
- Post-selection estimates and uncertainty can be very biased

# Consequences of ignoring model selection

- An example with predictors independent of response

# Model averaging

- Use all models (in the model set) and use the average value
- $\hat{\mu} = \sum_{S \in \mathcal{M}} c(S)\, \hat{\mu}_S$
- $\sum_{S \in \mathcal{M}} c(S) = 1$

- For parameters
- For predictions

# Model averaging

- How models should be weighted?
- AIC weights = $\frac{\exp(-\frac{1}{2}\Delta AIC_i)}{\sum_{model\ 1}^{model\ n}\exp(-\frac{1}{2}\Delta AIC_j)}$
- DIC weights
- BIC weights
- Posterior probabilities

# Model averaging

- An example
- Compare model averaging to estimator post selection

# Problems with model averaging

- Parameters must have the same interpretation in all models
  - Interactions
  - Linear and quadratic (or higher order) terms

# Problems with model averaging

- Not possible with models without a likelihood
- A simple average could then be used, but that ignores completely the fit data-model
- There are models for which the use of AIC (or other criteria) is still debated
  – (Generalized) Linear Mixed models

# Linear Mixed Models

- One-way Analysis of Variance model
- $Y_{ij} = \alpha_R + \alpha_i + e_{ij}$
- $e_{ij} independent\ and \sim Norm(0, sd = \sigma)$
- $\alpha_R$ and $\alpha_i$ are fixed values one is trying to estimate
- $Y_{ij} = \mu + \ a_i + e_{ij}$
- $a_i$ are random values, independent, and $\sim Norm(0, sd = \sigma_A)$

# Linear Mixed Models

- When to decide a factor should be fixed or random?
- If you want to estimate specific differences (e.g. between mowing regimes), it should be fixed
- If the name of levels has no meaning (sheep A, sheepB, etc.), then it is random
- If you are interested in the variability among levels, and that the levels can be considered as a sample from a wider population of levels, then it should be random

# Linear Mixed Models

- Changing a factor from fixed to random has important consequences for the statistical assumptions and properties of the model
- $Y_{ij} = \mu + \ a_i + e_{ij}$
- Independence and Normal distribution of $a_{ij}$ AND $e_{ij}$
- $Corr(Y_{i\ j}, Y_{i\ j'}) = \sigma_A^2 \big/ _{(\sigma_A^2 + \sigma_E^2)}$

# Linear Mixed Models and IC

- How to count parameters for the random effects?
- $a_i \sim N(0, \sigma_A) vs \left(a_1, a_2, \ldots, a_p\right): 1\ or\ p\ or\ ?$
- AIC in nlme or lmer calculated using 1

  (NOTE: Calculate AIC using ML, not REML!)
- No clear recommendation

# LMM and AIC

- Depends on the level of predictions: i or j («individuals» or «populations»)
- Population (i): 1 (variance)
- Individuals (j): CAIC, no implementation in R (as far as I know!)

*Biometrika* (2005), **92**, 2, *pp.* 351–370
© 2005 Biometrika Trust
*Printed in Great Britain*

**Conditional Akaike information for mixed-effects models**

BY FLORIN VAIDA

*Division of Biostatistics, Department of Family and Preventive Medicine, University of California at San Diego School of Medicine, La Jolla, California 92093, U.S.A.*

vaida@ucsd.edu

AND SUZETTE BLANCHARD

*Frontier Science and Technology Research Foundation Inc., Boston, Massachusetts 02215, U.S.A.*

# GLMM

- Logistic Regression Model:

$$logit(p_{ij}) = \mu + \ a_i + e_{ij}$$

- Random effects normally distributed on the logit scale
- Other parameterizations are possible

$$p_{ij} \sim Beta(\alpha, \beta)$$

Beta-binomial model

# GLMM

Beta distribution(s1,s2)