evropský sociální fond v ČR

EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tento projekt je spolufinancován Evropským sociálním fondem a Státním rozpočtem ČR
InoBio – CZ.1.07/2.2.00/28.0018

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness

INVESTMENTS IN EDUCATION DEVELOPMENT

Model Selection

Juliette Chamagne
Institute of Evolutionary Biology and Environmental Sciences
University of Zurich

Recap

- Linear Models (lm) assume:
  - Independence
  - Normality
  - Homogeneity
- Generalized Linear Models (glm) allow:
  - Linear predictor (Y ~ a + b*X_1 + c*X_2…)
  - Family distribution (variance)
  - Link function (mean)
- GLMs work with:
  - Continuous but positive (squewed) data (Gamma)
  - Count data (Poisson)
  - Proportion data or presence/absence data (binomial)

Model Selection

- Compare GLMs on the same data, with the same distributions, but with different link functions
- Compare GLMs with different predictors, to see which one(s) explain the data better

P Value Model Selection

- Run a nested sequence of models to successively compare and test terms in the models
- Omit non-significant terms in a process of model simplification
- Aim for a parsimonious Minimal Adequate (simplest) Model

Critique of Model Selection by hypothesis testing

- Subjective P-level
- Problem of multiple comparisons
- Problematic for non-nested models

Friedman's Paradox

- Even when the response is independent of the explanatory variables…
- When there are many explanatory variables (~50) variable selection methods will give high R^2s and many significant F and t values with coefficients biased away from zero.
- Partial resolution: Keep the number of candidate models small relative to the number of parameters to be estimated.

Model-selection uncertainty

- Remember that a given dataset is always a special case which may contain some unique (not general) effects that would not be present in replicated datasets of the same type.
- Avoid tailoring a model too much to a given dataset (over-fitting)

Maximum Likelihood

- **Given a set of data and a chosen a model** (we could try and compare several)...
- Maximum likelihood is a method for determining **which parameters of the model** produce the best model fit, as measured by **deviance**, and **make the data most likely to be observed**.
- No exact solutions but iterative approximations.
- The formula for the deviance changes for different types of data/error distributions
- However, for normal data the maximum likelihood estimate is the least squares estimate.

AIC

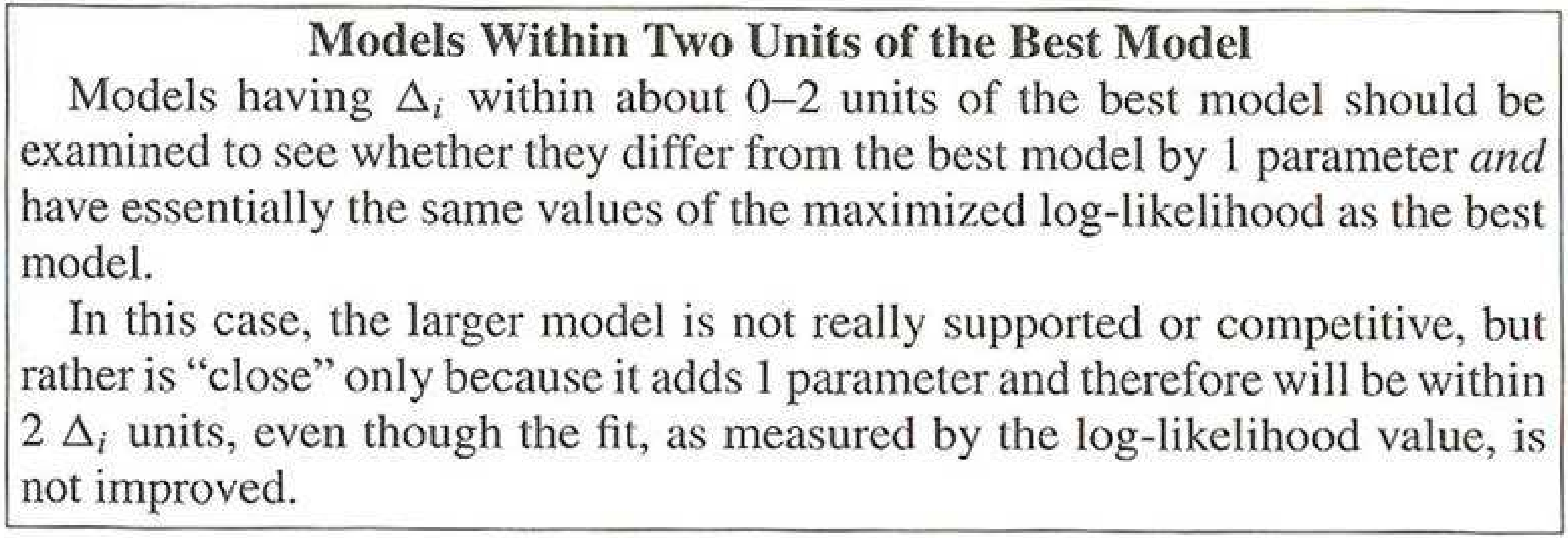
Decreases as more terms are added

$$\mathrm{AIC} = -2\ \ln[\mathcal{L}(\hat{\theta} \mid \mathbf{Y})] + 2K$$

Increases as more terms are added

- Trade-off between bias and variance (~complexity), or, under fitting and over fitting, that is fundamental to the principle of parsimony (Occam’s razor)
- Usually positive but can be negative, smaller values indicate better models
- Absolute value is of no interest due to relative scale that is also strongly dependent on sample size
- Recommended to report and compare AIC differences

Comparing models by AIC differences



Models Within Two Units of the Best Model

Models having Δ_i within about 0–2 units of the best model should be examined to see whether they differ from the best model by 1 parameter *and* have essentially the same values of the maximized log-likelihood as the best model.

In this case, the larger model is not really supported or competitive, but rather is "close" only because it adds 1 parameter and therefore will be within 2 Δ_i units, even though the fit, as measured by the log-likelihood value, is not improved.

Advantages of AIC

- AIC can be used to compare non-nested models while likelihood ratio tests cannot.
- Allows ranking of models
- Allows ratio of evidence for different models
- Allows multimodel inference using parameter weighted averages
- Order of calculating AICs for different models does not matter

Limits on Model Comparison

- AIC can only be used to compare different models applied to **exactly** the same dataset.
- Different transformations cannot be used when comparing models using AIC.
- Instead use GLMs to compare models with different link functions.
- But, is only straight-forward when using the same error distribution in the GLM (see B&A p.318; Faraway 2006 p. 138).
- Cannot know how close to the ‘true’ model candidate models are (even if truth is assumed to exist), only their relative rankings

Pitfalls

- Large numbers of models for small datasets
- Bad candidate models and subset models
- Models within 2 IT units are approximately equally good but may contain 1 useless parameter (since the penalty term is 2p)
- Use of AIC with small samples
- AIC 'best' models may contain parameters with little support (estimates close to zero etc.).

Important things to keep in mind

- Models have to make sense
- Choosing the best model between two bad models is still bad
- Don’t try to fit all possible models. Select a few that correspond to the hypotheses you want to test.

Model selection in R: P value

```
> mod1 <- glm(Y ~ X1 + X2, data)
> mod2 <- glm(Y ~ X1 + X2 + X1:X2, data)
> anova(mod1, mod2)
    # gives a p-value for how different the two models are
    # if the difference is significant, take the model with the
smallest deviance (or the residual SS in case of lm)
    # if there is no difference, take the simpler model
    # mod1 is “nested” within mod2
```

Model selection in R: AIC

```
> mod1 <- glm(Y ~ X1 + X2, data)
> mod2 <- glm(Y ~ X1 + X3 + X4 + X1:X3, data)
> AIC(mod1, mod2)
    # the lower the AIC, the better model
    # mod1 and mod2 don’t have to be nested
    # mod1 and mod2 should still be working with the same
dataset, and the same Y
```

Example in R: the forest dataset

```
> mod1 <- lm(Prod ~ SpDiv, data=forest)
> mod2 <- lm(Prod ~ SpDiv + ForType, data=forest)
> anova(mod1, mod2)
Analysis of Variance Table

Model 1: Prod ~ SpDiv
Model 2: Prod ~ SpDiv + ForType
  Res.Df  RSS       Df     Sum of Sq    F    Pr(>F)
1  103    6.3835
2   101    2.0374    2  4.346      107.72    < 2.2e-16 ***
```

# the two models are significantly different, so take the one with the smallest Residual SS: it means that it has less unexplained variance.

Example in R: the forest dataset

```
> mod1 <- lm(Prod ~ SpDiv, data=forest)
> mod2 <- lm(Prod ~ SpDiv + ForType, data=forest)
> mod3 <- lm(Prod ~ SpDiv * ForType, data=forest)
> mod4 <- lm(Prod ~ SpDiv, data=forest)
> AIC(mod1, mod2, mod3, mod4)
     df         AIC
mod1  3    9.950882
mod2  5 -105.961317
mod3  7 -133.809405
mod4  3    9.950882
# same result with the AIC criterion: mod3 has a better AIC (lower)
# here all models don't have to be nested (mod4 is not nested in mod1)
```

Exercises in R

- For every dataset you’ve been working on this week (GLMs), use the AIC() to compare the different models.
- Does the result of the AIC make sense? Compare with the diagnostic plots, i.e. plot(model), and with what you see graphically, i.e. qplot(X,Y, geom=c(“point”,”smooth”), method(link))