

22S:138 Model Comparison

Lecture 18a
Nov. 12, 2004

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

Model comparison for nested vs. non-nested models

- nested models: two regression-type models in which the predictors in the smaller model are a subset of the predictors in a larger model
 - larger model will fit better but will be more difficult to fit and to interpret
 - key questions in model comparison
 - * is improvement in fit substantial enough to justify increased difficulty in fitting and interpreting
 - * are priors on additional parameters reasonable
- non-nested models
 - different link functions in GLMs
 - non-nested sets of predictors

Model comparison

- often there are several plausible candidate models
 - different candidate predictor variables in regression
 - different link functions in generalized linear models
 - different assumptions regarding form of likelihood
 - different priors
- statisticians often will compare the fit of several models in order to choose the “best” one
 - then assess whether that one is adequate
- alternative: Bayesian model-mixing
 - does prediction using a weighted combination of all candidate models

Tools for Bayesian model comparison

- Bayes factors and approximations to them
- Deviance Information Criterion

Frequentist use of deviance as measure of model fit in linear and generalized linear models

Example:

Dataset is counts of how many beetles were killed r_i , $i = 1, \dots, 8$ in 8 groups of beetles exposed to different doses of an insecticide. Each group i had n_i beetles in it.

- consider a “saturated model” for a particular dataset
 - has a parameter for every observation in the dataset so its fit is “perfect”
 - not useful, since it is no simpler than the entire original dataset
 - but it provides a benchmark to which to compare the fit of other models

7

against the general alternative

- under certain conditions, deviance has an asymptotic chi-square distribution with degrees of freedom equal to the difference between the number of parameters in the saturated model and the number of parameters in the model being evaluated

- saturated model for beetles data would have 8 parameters: p_i , $i = 1, \dots, 8$, the population proportion killed at each of the 8 dose levels
- the frequentist point estimate of each p_i would be $\frac{r_i}{n_i}$

- now consider a more useful model that lets us quantify the dose-response

$$\text{logit}(p_i) = \alpha + \beta(x_i - \bar{x})$$

- has only 2 parameters
- will not fit the data as perfectly as the saturated model

- notation: let $\log L(\hat{\theta}; \mathbf{y})$ denote the maximum of the log likelihood for a particular model

- *deviance* in GLM is defined as

$$-2 [\log L(\hat{\theta}_{\text{model of interest}}; \mathbf{y}) - \log L(\hat{\theta}_{\text{saturated}}; \mathbf{y})]$$

- this is the likelihood-ratio statistic for testing the null hypothesis that the model holds

8

Frequentist deviance for models for beetles data

```
fit.beetles(beetles)
      [,1] [,2]
[1,]    6  53
[2,]   13  47
[3,]   18  44
[4,]   28  28
[5,]   52  11
[6,]   52   7
[7,]   61   1
[8,]   60   0

Call:
glm(formula = respmat ~ beetles$V1, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5213 -0.6270  0.8705  1.2575  1.6487

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -59.869      5.100  -11.74  <2e-16 ***
beetles$V1    33.784      2.866   11.79  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 280.866  on 7  degrees of freedom
Residual deviance: 11.474  on 6  degrees of freedom
AIC: 41.803
```

```
Call:
glm(formula = respmat ~ beetles$V1, family = binomial(link = probit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4994  -0.6939   0.7942   1.1473   1.3076

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -34.501     2.616  -13.19  <2e-16 ***
beetles$V1    19.478     1.469   13.26  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 280.866  on 7  degrees of freedom
Residual deviance:  10.368  on 6  degrees of freedom
AIC: 40.698
```

```
Call:
glm(formula = respmat ~ beetles$V1, family = binomial(link = cloglog))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7906  -0.6252   0.0838   0.4158   1.4120

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -39.035     3.182  -12.27  <2e-16 ***
beetles$V1    21.733     1.766   12.31  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 280.8664  on 7  degrees of freedom
Residual deviance:   4.0124  on 6  degrees of freedom
AIC: 34.342
```

Complementary log-log link:

$$cloglog(p) = \log(-\log(1 - p))$$

Deviance Information Criterion

- Spiegelhalter D J, Best N G, Carlin B P and van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc. B.* 64, 583-640.
- to compare fit and predictive ability of Bayesian models
- penalty for model complexity
- also provides estimate of number of free parameters in the model
 - highly correlated parameters and parameters that are strongly influenced by their priors count for less than 1 each
 - called the *effective number of parameters*
- built into WinBUGS
- can be used to compare non-nested models

- but response variable must have same form in all models
 - e.g. you couldn't use it to compare two regression models, one with y's untransformed and one with y's log transformed
- uses a version of the deviance from which the log likelihood of the saturated model is *not* subtracted off
- let $D(\mathbf{y}, \boldsymbol{\theta}) = -2\log p(\mathbf{y}|\boldsymbol{\theta})$
- we want two quantities, which can be approximated using MCMC sampler output
 - $\hat{D}_{avg}(\mathbf{y})$: D averaged over the posterior distribution of $\boldsymbol{\theta}$
 - $D_{\hat{\boldsymbol{\theta}}}(\mathbf{y})$: D evaluated at the posterior mean of $\boldsymbol{\theta}$
- then the effective number of parameters is estimated as

$$p_D = \hat{D}_{avg}(\mathbf{y}) - D_{\hat{\boldsymbol{\theta}}}(\mathbf{y})$$

- and the DIC is

$$\begin{aligned} DIC &= \hat{D}_{avg}(\mathbf{y}) + p_D \\ &= 2\hat{D}_{avg}(\mathbf{y}) - D_{\hat{\boldsymbol{\theta}}}(\mathbf{y}) \end{aligned}$$

- DIC is an approximation to the expected predictive deviance and has been suggested as an indicator of model fit when the goal is to pick a model with the best out-of-sample predictive ability
- smaller values of DIC suggest better model fit