Model Checking and Improvement III

Statistics 220

Spring 2005

Copyright ©2005 by Mark E. Irwin

Model Comparison

There are two situations where comparing models may be reasonable

1. Nested models:

   Example:

Model 1	Model 2
$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$	$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

   The first model is a special case of the second model ($\beta_2 = 0$).

   In most cases the larger model will fit the data better but can be more difficult to interpret and compute. Two questions of interest in this comparison are

(a) Is the improvement in fit large enough to justify the additional difficulty in interpretation and computation?
(b) Is the prior distribution on the additional parameters reasonable?

Note: This second question is why I noted that the larger model will usually fit better, instead of always. A bad prior may bias the fits for small sample sizes. In a likelihood based analysis, the larger model must always give a better fit (assuming deviance is the measure of fit).

Standard hypothesis testing methods address the first question in frequentist analyses. For example, the use of an F-test to compare two linear regression models.

The approach we will take here is one that measures the distance of the data to each of the models.

- θ: parameters in first model
- ϕ: additional parameters in second model

So we want to compare $p(\theta|y)$ and $p(y^{rep}|y)$ with $p(\theta, \phi|y)$ and $p(y^{rep}|y)$

2. Non-nested models:

   Example:

$$\begin{array}{cc} \text{Model 1} & \text{Model 2} \\ y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i & y_i = \beta_0 + \beta_2 x_{i2} + \epsilon_i \end{array}$$

   In this case, you can't make one model a special case of the following.

   Comparisons of this sort can be useful, if for example, you wish to compare two regression with different sets predictors to see which one does better. Maybe for cost considerations you can only afford one predictor, so you need to figure out which is the best.

Expected Deviance

Last class we discussed the use of

$$T(y,\theta) = \sum_i \frac{(y_i - E[y_i|\theta_i])^2}{\mathrm{Var}(y_i|\theta_i)}$$

as measure of model fit.

Another option, which tends to work better for our purposes is the deviance

$$D(y,\theta) = -2\log p(y|\theta)$$

This has ties to the Kullback-Leibler information

$$\begin{aligned} H(\theta) &= \int \log\left(\frac{f(y)}{p(y|\theta)}\right) f(y)dy \\ &= \int \log f(y) f(y)dy - \int \log p(y|\theta) f(y)dy \end{aligned}$$

So

$$2H(\theta) = \int D(y,\theta) f(y) dy + 2 \int \log f(y) f(y) dy$$

which implies

$$E[D(y,\theta)] = 2H(\theta) - 2 \int \log f(y) f(y) dy$$

Thus the expected deviance (averaged over the true sampling distribution $f(y)$) is twice $H(\theta)$ minus a factor that doesn't depend on θ.

As we have discussed before, as the sample size goes to infinity, our posterior inference goes to the model with the smallest value of $H(\theta)$. So this suggests using an estimate of the expected deviance as a measure of overall model fit.

In using the deviance to measure model fit there are two approaches.

1. Plug in estimate of θ

$$D_{\hat{\theta}}(y) = D(y, \hat{\theta}(y))$$

where $\hat{\theta}(y)$ is an estimate of θ based on the data y, say, for example, the posterior mean of θ.

2. Average over the posterior realizations of θ

$$D_{avg}(y) = E[D(y, \theta)|y]$$

which can be estimated using posterior simulations $\theta^1, \ldots, \theta^L$ by

$$\hat{D}_{avg}(y) = \frac{1}{L} \sum_{l=1}^{L} D(y, \theta^l)$$

$\hat{D}_{avg}(y)$ is a better measure of model error since it averages over our uncertainty about θ.

$D_{\hat{\theta}}(y)$ tends to indicate a better fit than we really have as it calculates the discrepancy under a more probable θ

Counting Parameters and Model Complexity

While $D_{\hat{\theta}}(y)$ is not a good measure of model fit, it is an interesting descriptor as

$$p_D^{(1)} = \hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$$

is a measure of the effective number of parameters in the Bayesian Model.

An alternative measure is

$$p_D^{(2)} = \frac{1}{2}\widehat{\mathrm{Var}}(D(y,\theta)|y) = \frac{1}{2}\frac{1}{L-1}\sum_{l=1}^{L}(D(y,\theta^l) - \hat{D}_{avg}(y))^2$$

Both of these measures are based on the distribution of $D(y,\theta)$, relative to its minimum, having an asymptotic χ^2 distribution. (Note $E[\chi^2_\nu] = \nu$ and $\mathrm{Var}(\chi^2_\nu) = 2\nu$.)

The second of these ($p_D^{(2)}$) is the preferred as the asymptotics should work a bit better (less worry about bias).

The way to think of p_D is a the number of 'unconstrained' parameters in the model. A parameter will count as

- 1 if it is completely unconstrained (no information about it in the prior)
- 0 if it is completely constrained (all information about if is in the prior)
- intermediate if information about it comes from the data and the prior

Deviance Information Criterion

A useful measure for model selection is based on

$$D_{avg}^{pred}(y) = E[D(y^{rep}, \hat{\theta}(y))]$$

where $D(y^{rep}, \hat{\theta}(y)) = -2 \log p(y^{rep}|\theta)$

This can be estimated by the *Deviance Information Criterion* (DIC)

$$DIC = \hat{D}_{avg}^{pred}(y) = 2\hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$$

This can be thought of as

$$DIC = \hat{D}_{avg}(y) + p_D^{(1)}$$

the average deviance plus a penalty term, giving it the same flavour as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) used in likelihood analysis

$$AIC = -2 \log p(y|\hat{\theta}) + 2K$$

$$BIC = -2 \log p(y|\hat{\theta}) + K \log n$$

where K is the number of parameters in the model

Note that DIC is different than many of the other measures discussed in that it uses as estimated value for θ instead of averaging the posterior distribution of $\theta|y$.

The goal of DIC is use it to predict a model with the best out-of-sample predictive power. The book motivates DIC by starting with

$$D_{avg}^{pred}(y) = E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^{rep} - E[y_i^{rep}|y])^2\right]$$

(This is for normal likelihoods, though the book doesn't mention it.)

Example: Rat Tumor Rates

Model	$D_{\hat{\theta}}$	$\hat{D}_{avg}$	$p_D^{(2)}$	DIC
Shrinkage	167.9	253.0	85.1	338.1
Common Theta	343.8	344.8	1	345.8

(Calculated by WinBugs)

So this suggests that the Shrinkage model is a preferable model to the Common Theta model (as we have already seen).

```
          mean      sd  2.5%   25%   50%   75% 97.5% Rhat n.eff
theta      0.2     0.0   0.1   0.1   0.2   0.2   0.2    1  5000
a        139.7  5659.1   0.0   0.0   0.3   1.1  58.2    1  3400
b        733.1 29395.6   0.0   0.1   0.4   3.0 328.5    1  3100
u          0.4     0.2   0.1   0.2   0.3   0.5   0.9    1  5000
v          5.2    28.6   0.1   0.5   1.1   2.8  28.1    1  3100
deviance 344.8     1.4 343.8 343.9 344.3 345.1 348.8    1  5000
 pD = 1 and DIC = 345.8 (using the rule, pD = var(deviance)/2)
```

Comment on WinBugs output and Table 6.2 in the text

In WinBugs, as noted in the output, the reported $p_D = p_D^{(2)}$ and $DIC = \hat{D}_{avg} + p_D^{(2)}$, not $DIC = \hat{D}_{avg} + p_D^{(1)}$ as you get if you follow the development of the math in the text.

In the table on the previous page, $D_{\hat{\theta}} = \hat{D}_{avg} - p_D^{(2)}$ which doesn't quite match the books development, which has $D_{\hat{\theta}} = \hat{D}_{avg} - p_D^{(1)}$.

I suspect that Table 6.2 in the text, which is comparison of 3 Normal random effects models (SAT coaching example) was calculated by WinBugs as well as the same relationships between $D_{\hat{\theta}}, \hat{D}_{avg}, p_D^{(2)}$, and DIC hold.

Note that these deviations from the books development shouldn't be very important as sample sizes increase $p_D^{(1)}$ and $p_D^{(2)}$ should approach each other.

Comment of p_D

As mentioned earlier, p_D can be thought of as a measure of the effective number of unconstrained parameters in a model. Note that estimates of this quantity don't have to be less than the number of actual parameters in the model.

For example, for the rat tumor example with the shrinkage model, $p_D^{(2)} = 85.1$. However the actually number of parameters in the model is 73 (71 θs, α, and β). I suspect (though I'm not sure) that this difference is due to randomness in the data.

An analogue would trying to estimate the degrees of freedom in a χ^2 test based on the observed test statistics and an assumption that the data was really generated under the null hypothesis.

Bayes Factors

As discussed earlier, Bayes theorem can be thought of in terms of posterior odds, which satisfies

$$\text{Posterior Odds} = \text{Prior Odds} \times \text{Likelihood Ratio}$$

Thus the posterior odds of two models H_1 and H_2 is given by

$$\frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2)}{p(H_1)} \times \frac{p(y|H_2)}{p(y|H_1)}$$

where

$$p(y|H_i) = \int p(\theta_i|H_i)p(y|\theta_i, H_i)d\theta_i$$

The ratio of the marginal likelihoods

$$BF(H_2; H_1) = \frac{p(y|H_2)}{p(y|H_1)}$$

is known as the Bayes Factor. Its a measure of how much the data favours model H_2 over H_1

Note that the Bayes factor is only defined when the marginal likelihood of each model is proper.

Consider the model

$$\begin{aligned} y|\theta &\sim N(\theta, 1) \\ p(\theta) &\propto 1 \end{aligned}$$

Then

$$p(y) \propto \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\theta)^2\right) d\theta = 1$$

Thus marginally y is uniform on $(-\infty, \infty)$ and thus has an improper distribution.

(Aside: even though the marginal density on y is improper, the posterior $p(\theta|y)$ is proper as we have seen. It happens that things cancel properly when going to the posterior.)

Note that the Bayes factor only really make sense when there are a discrete number of models being investigated.

A good example of where Bayes factor have been used is in genetic counselling, such as was done for diseases like Huntington's (an autosomal dominant trait with the gene on chromosome 4) in the late 80's and early 90's.

$H_1 =$ subject is not at risk for disease (not a gene carrier)

$H_2 =$ subject is at risk (gene carrier)

$y =$ marker data on subject and relatives and disease status in relatives (subject is currently unaffected).

Of interest is $P[H_2|y]$ or equivalently $\frac{p(H_2|y)}{p(H_1|y)}$.

As testing was usually done in people with one affected parent, lets assume that $\frac{p(H_2)}{p(H_1)} = 1$. It will be different if other affection pattern in the family led to the test. Both parents being affected would raise the prior odds to 3.

The likelihoods under the two situations was calculated using a peeling algorithm.

With these prior odds, the posterior odds is actually the Bayes factor. I believe with the standard test of the time, the largest the Bayes factor could be is about 33 and the smallest is about $\frac{1}{33}$ (the marker used was about 3cM from the gene I think).

Note current technology eliminates the needs for these sorts of calculations for many diseases. Though not completely. Even though the Huntington gene (IT-15) has been found, sequenced, and is somewhat understood, the recent test for this condition isn't perfect. Thus this sort of analysis (though) with different sorts of calculations due to the different types of data.

While Bayes factors can be calculated for models described by continuous parameter, then tend not to be useful.

For example consider the set of normal based models H_τ

$$\begin{aligned} y_{ij}|\theta &\overset{ind}{\sim} N(\theta_j, \sigma^2) \\ \theta_j &\overset{iid}{\sim} N(\mu, \tau^2) \\ \mu &\sim p(\mu) \qquad \text{(Proper)} \end{aligned}$$

While the Bayes factor is well defined in this case

$$BF(H_\tau; H_0) = \frac{p(y|\tau)}{p(y|\tau = 0)}$$

it not particularly helpful for picking which H_τ is the best model.

Also in these situations, the Bayes factor can break down and become unstable, particularly when $p(\mu)$ is pushed to the limit to approximate an improper prior.