Biostatistics 682: Applied Bayesian Inference Lecture 14: Model selection and goodness-of-fit checks

Jian Kang

Department of Biostatistics
University of Michigan, Ann Arbor

Objectives

- We will study methods for model comparisons and checking for model adequacy
- For model comparisons there are a finite number of candidate models and we want to select one
  - Bayes factors
  - Cross validation
  - Deviance information criteria (DIC)
- In cases where multiple models fit well, we can consider
  - Bayesian model averaging (BMA)
- After selecting a model, we want to test whether it fits the data well
  - Posterior predictive checks

Quote from the FDA

- FDA recommends you investigate all assumptions important to your analysis.
- You may summarize this comparison using a Bayesian p-value (Gelman et al., 1996, 2004), the predictive probability that a statistic is equal to or more extreme than that observed under the assumptions of the model.
- You may also assess model checking and fit by Bayesian deviance measures, such as the Deviance Information Criterion as described in Spiegelhalter et al. (2002).
- Alternatively, two models may be compared using Bayes factors.

Bayes factors (BF)

- Consider two models: $\mathcal{M}_1$ and $\mathcal{M}_2$
- For example, $Y \sim \text{Binomial}(n, \theta)$ and the two models are

$$\mathcal{M}_1 : \theta = 0.5 \qquad \text{and} \qquad \mathcal{M}_2 : \theta \neq 0.5$$

- Another example, $Y_1, Y_2, \ldots, Y_n$ is a time series and

$$\mathcal{M}_1 : \text{Cor}(Y_{t+1}, Y_t) = 0, \qquad \text{and} \qquad \mathcal{M}_2 : \text{Cor}(Y_{t+1}, Y_t) > 0$$

- Another example,

$$\mathcal{M}_1 : \text{E}(Y) = \beta_0 + \beta_1 X, \qquad \text{and} \qquad \mathcal{M}_2 : \text{E}(Y) = \beta_0 + \beta_1 X + \beta_2 X^2.$$

Bayes factors (BF)

- This is similar to hypothesis testing
- As before we proceed by computing the posterior probabilities of the two models
- This requires prior probabilities $\pi(\mathcal{M}_1)$ and $\pi(\mathcal{M}_2)$ on the model
- This is different than the priors for parameters
- We can make the probabilistic statement that the "with the prior knowledge, the quadratic model is five times more likely than a linear model"

Bayesian factors (BF)

- The Bayes factor for model 2 compared to model 1 is

$$BF = \frac{\text{Posterior odds}}{\text{Prior odds}} = \frac{\pi(\mathcal{M}_2 \mid Y)/\pi(\mathcal{M}_2 \mid Y)}{\pi(\mathcal{M}_2)/\pi(\mathcal{M}_1)} = \frac{\pi(Y \mid \mathcal{M}_2)}{\pi(Y \mid \mathcal{M}_1)}.$$

- Rule of thumb: BF > 10 is strong evidence for $\mathcal{M}_2$
- Rule of thumb: BF > 100 is decisive evidence for $\mathcal{M}_2$
- In linear regression, BIC approximates the BF comparing a model to the null model

Example

- $Y \sim \text{Binomial}(n, \theta)$ with

$$\mathcal{M}_1 : \theta = \theta_0, \qquad \mathcal{M}_2 : \theta \neq \theta_0$$

- $\pi(Y \mid \mathcal{M}_1)$ is just the binomial density with $\theta = \theta_0$.

$$\pi(Y \mid \mathcal{M}_1) = \binom{n}{Y} \theta_0^Y (1 - \theta_0)^{n-Y}$$

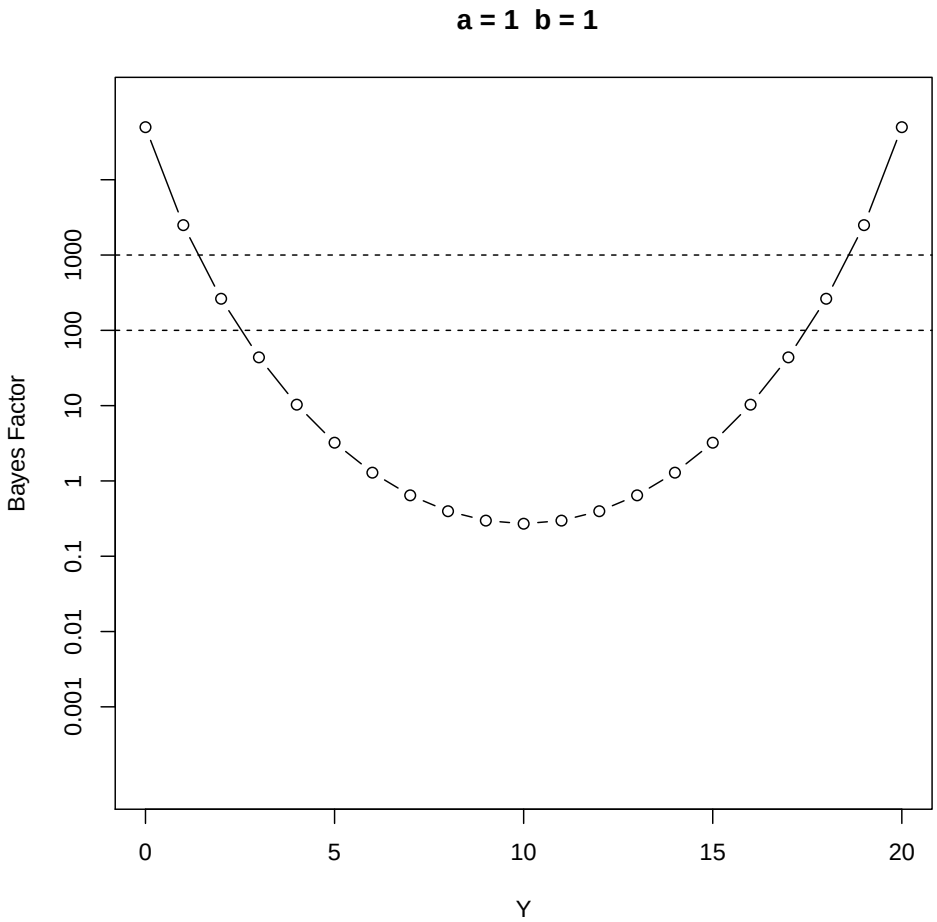
- M_2 involves an unknown parameter θ.
- This requires a prior, say $\theta \sim \text{Beta}(a, b)$, and integration

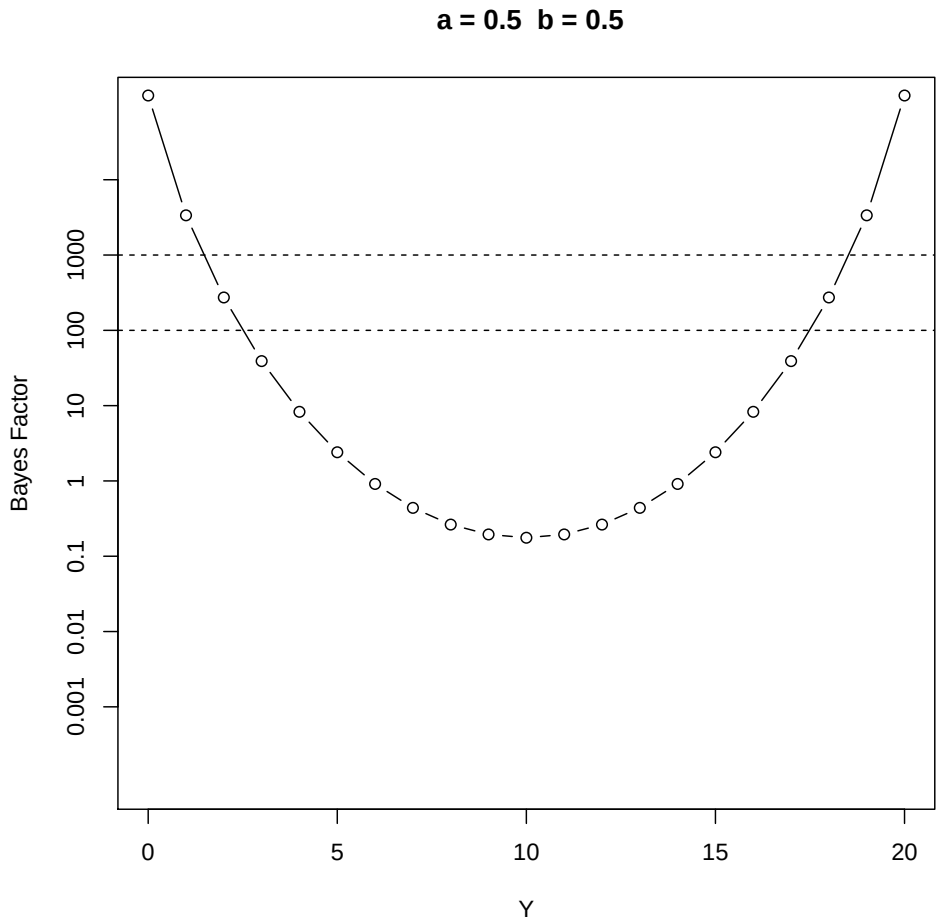
$$\pi(Y \mid \mathcal{M}_2) = \int \pi(Y, \theta) d\theta = \binom{n}{Y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(Y+a)\Gamma(n-Y+b)}{\Gamma(n+a+b)}.$$

- The BF is

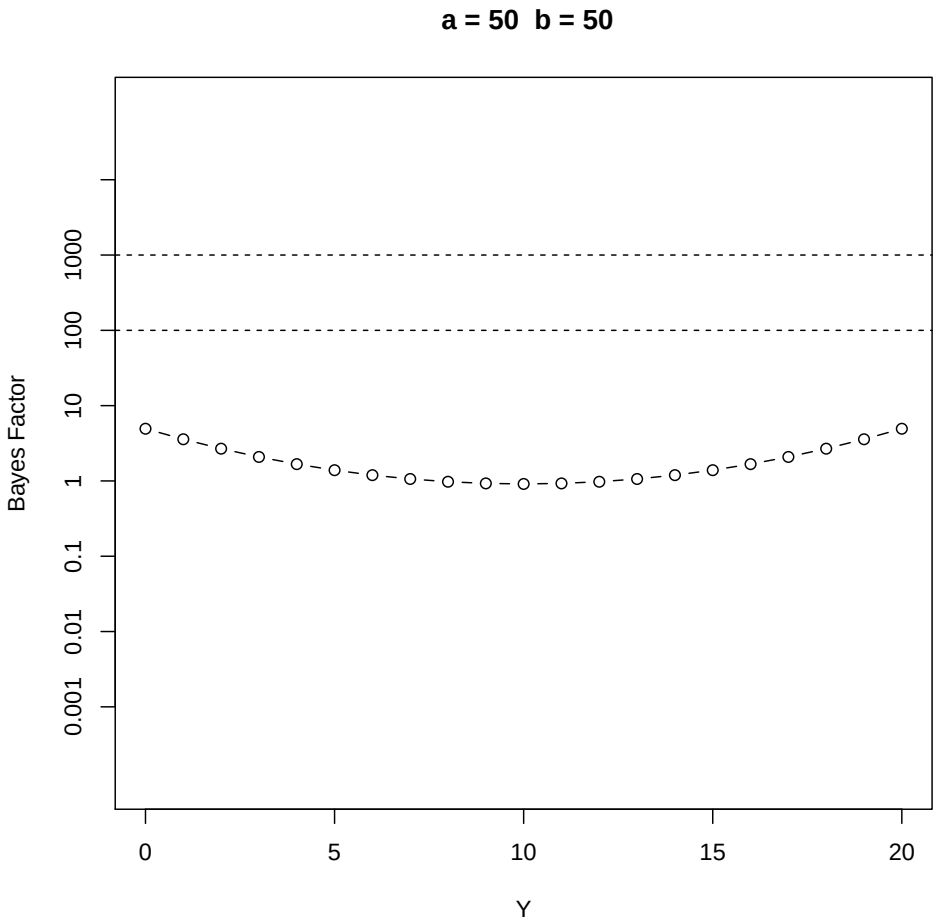
$$\text{BF}(\mathcal{M}_2 \mid \mathcal{M}_1) = \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(Y+a)\Gamma(n-Y+b)}{\Gamma(n+a+b)}}{\theta_0^Y (1 - \theta_0)^{n-Y}}$$

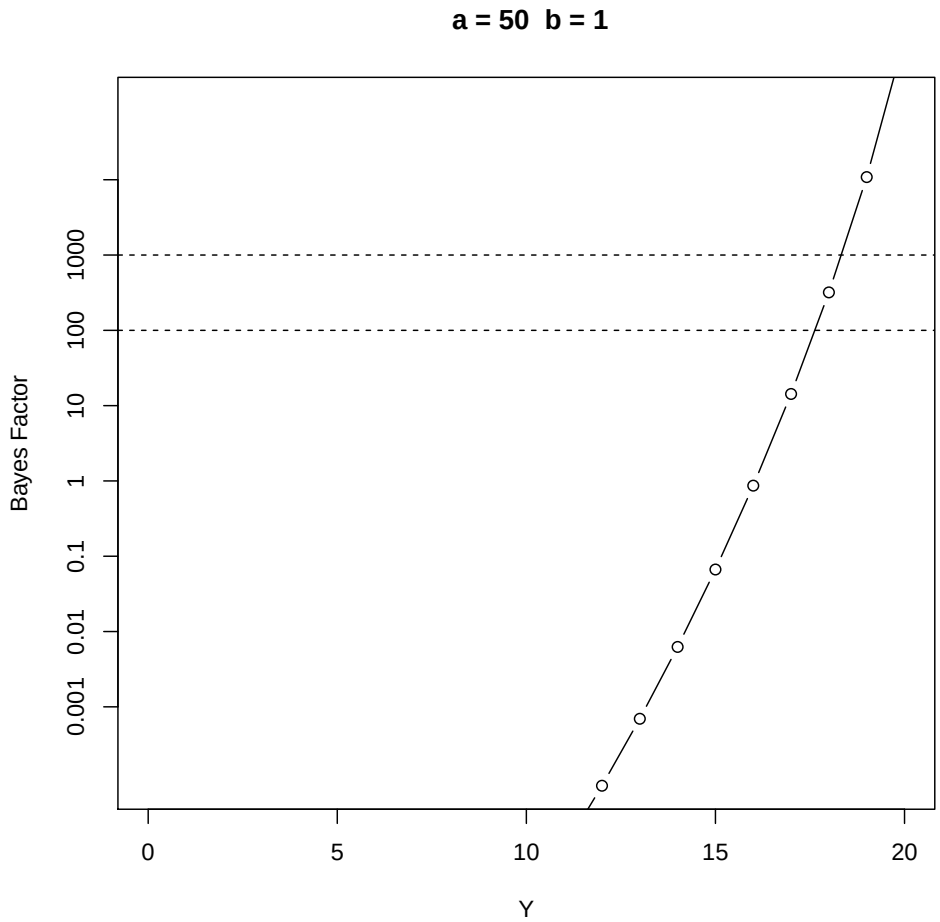
Example: $\theta_0 = 0.5$ and $n = 20$





Example: $\theta_0 = 0.5$ and $n = 20$





Limitations with Bayes factors

- Often required intractable integral computation for complex model
  - Monte Carlo method?
- Requires proper prior specifications
- Can be sensitive to prior specifications (Lindley's paradox)
- In some scenarios, we can consider Bayesian model averaging

Bayesian model averaging

- Consider the linear regression example

$$\mathcal{M}_1 : \mathrm{E}(Y) = \beta_0 + \beta_1 X \text{ v.s. } \mathcal{M}_2 : \mathrm{E}(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$$

- Say we have fit both models and found that both are about equally likely, but that $\mathcal{M}_1$ is slightly preferred
- For prediction, $\hat{Y}$, we could simply take the prediction that comes from fitting $\mathcal{M}_1$
- But the prediction from $\mathcal{M}_2$ is likely different and nearly as accurate
- Also, taking the prediction from $\mathcal{M}_1$ suppresses our uncertainty about the form of the model

Bayesian model averaging

- Let $\hat{Y}_k$ be the prediction from model $\mathcal{M}_k$ for $k = 1, 2$
- The model averaged predictor is

$$\hat{Y} = w\hat{Y}_1 + (1 - w)\hat{Y}_2$$

- It can be shown that the optimal weight w is the posterior probability of $\mathcal{M}_1$.
- Averaging adds stability
- In linear regression with p predictors the prediction is a weighted average of 2^p possible models
- We can implement this by introducing latent indicators.

Cross validation

- Another very common approach is cross validation
- This is exactly the same procedure used in classical statistics
- This operates under the assumption that the "true" model likely produces better out-of-sample prediction than competing models
- Pros: conceptually simple, intuitive, and broadly applicable
- Cons:
  - Slow because it requires several model fits (can we do better?)
  - it is hard to say a difference is statistically significant.

K-fold cross validation

Step 0: Split the data into K equally-sized groups

Step 1: Set aside group k as test set and fit the model to the remaining $K-1$ groups

Step 2: Make predictions for the test set k based on the model fit to the training data

Step 3: Repeat steps 1 and 2 for $k = 1, \ldots, K$ giving a predicted value $\hat{Y}_i$ for all n observations

Step 4: Measure prediction accuracy, e.g.,

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Variants

- Usually K is either 5 or 10
- $K = n$ is called "leave-one-out" cross-validation, which is great but slow
- The predictive value $\hat{Y}_i$ can be either the posterior predictive mean or median
- Mean squared error (MSE) can be replaced with Mean absolute deviation

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \hat{Y}_i|.$$

Deviance information criteria (DIC)

- DIC is a popular Bayesian analog of AIC and BIC
- Unlike CV, DIC requires only one model fit to the entire dataset
- Unlike BF, it can be applied to complex models
- However, proceed with caution
- DIC really only applies when the posterior is approximately normal, and will give misleading results when the posterior far from normality, e.g. bimodal
- DIC is also criticized for selecting overly-complex models

Deviance information criteria (DIC)

- Recall $\mathrm{DIC} = \bar{D} + p_D$ where
  - $\bar{D} = \mathrm{E}(-2 \log \pi(Y \mid \theta))$ is the posterior mean of the deviance
  - p_D is the effective number of parameters
- Models with small $\bar{D}$ fit the data well
- Models with small p_D are simple
- We prefer models that are simple and fit well, so we select the model with the smallest DIC

DIC Discussion

- The effective number of parameters is a useful measure of model complexity
- Intuitively, if there are p parameters and we have uninformative priors then $p_D \approx p$
- However, $p_D << p$ if there are strong priors
- For example, how many free degrees of freedom do we have with $\theta \sim \text{Beta}(1, 1)$ versus $\theta \sim \text{Beta}(1000, 1000)$

DIC Discussion

- As with AIC or BIC, we compute DIC for all models under consideration and select the one with smallest DIC
- Rule of thumb:a difference of DIC of less than 5 is not definitive and a difference greater than 10 is substantial
- As with AIC or BIC, the actual value is meaningless, only differences are relevant

Posterior predictive checks

- After comparing a few models, we settle on the one that seems to fit the best
- Given this model, we then verify it is adequate
- The usual residual checks are appropriate here: qq-plots
- A uniquely Bayesian diagnostic is the posterior predictive check
- This leads to the Bayesian p-value

Posterior predictive distributions

- Before discussing posterior predictive checks, let's review Bayesian prediction in general
- The plug-in approach would fix the parameters θ at the posterior mean $\hat{\theta}$ and the predict $y_{\text{new}} \sim f(y \mid \hat{\theta})$
- This suppresses uncertainty in θ
- We would like to propagate this uncertainty through to the prediction

Posterior predictive distributions (PPD)

- We really want to PPD

$$\pi(y_{\text{new}} \mid Y) = \int \pi(y_{\text{new}}, \theta \mid Y) d\theta = \int \pi(y_{\text{new}} \mid \theta) \pi(\theta \mid Y) d\theta.$$

- MCMC easily produces draws from this distribution
- To make S draws from the PPD, for each of the S MCMC draws of θ we draw a y_{new}.
- This gives draws from the PPD and clearly accounts for uncertainty in θ.

Posterior predictive checks

- Posterior predictive checks sample many data sets from the PPD with the identical design (same n, same X) as the original dataset
- We then define a statistic describing the dataset, e.g.,

$$d(Y) = \max\{Y_1, \ldots, Y_n\}$$

- Denote by d_0 the statistic for the original data set and by d_s the statistic from the simulated data set s.
- If the model is correct, then d_0 should fall in the middle of the $d_1, \ldots, d_S$.

Posterior predictive checks

- A measure of how extreme the observed data is relative to this sampling distribution is the Bayesian p-value

$$p = \sum_{s=1}^{S} I(d_s > d_0)$$

- If p is near zero or one the model does not fit