

Chapter 5

Model comparisons

Model selection

- ▶ We now have many potential models in our arsenal
- ▶ For a given dataset, how do determine whether a simple model is sufficient or if we need to bring out the “big guns”?
- ▶ Is there a “right” model? Probably not
- ▶ A **statistical model** is a mathematical representation of the system that includes errors and biases in the observation process
- ▶ All models are simplifications of reality
- ▶ We want a model that is as simple as possible yet seems to fit the data reasonably well

Outline

These notes cover Chapter 5

- ▶ For model comparisons there are a finite number of candidate models and we want to select one
 - ▶ Bayes factors
 - ▶ Stochastic search variable selection
 - ▶ Cross validation
 - ▶ Deviance information criteria (DIC)
 - ▶ Watanabe-Akaike information criteria (WAIC)
- ▶ In cases where multiple models fit well, we will discuss
 - ▶ Bayesian model averaging (BMA)
- ▶ After selecting a model, we want to test whether it fits the data well
 - ▶ Posterior predictive checks

Bayes factors (BF)

- ▶ In some sense BFs are the gold standard
- ▶ Say we are comparing two models, \mathcal{M}_1 and \mathcal{M}_2
- ▶ For example, $Y \sim \text{Binomial}(n, \theta)$ and the two models are

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

- ▶ Another example, Y_1, Y_2, \dots, Y_n is a time series and

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

- ▶ Another example,

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

Bayes factors (BF)

- ▶ This is really the same as hypothesis testing, and in fact Bayes factors are the gold standard for hypothesis testing
- ▶ As before we proceed by computing the posterior probability of the two models
- ▶ This requires prior probabilities $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$
- ▶ This is not a prior on a parameter, it is a prior on the model!
- ▶ This approach permits statements such as “Given the data we have observed, the quadratic model is 5 times more likely than a linear model”

Bayes factors (BF)

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

$$BF = \frac{\text{Posterior odds}}{\text{Prior odds}} = \frac{p(\mathcal{M}_2|\mathbf{Y})/p(\mathcal{M}_1|\mathbf{Y})}{p(\mathcal{M}_2)/p(\mathcal{M}_1)} = \frac{p(\mathbf{Y}|\mathcal{M}_2)}{p(\mathbf{Y}|\mathcal{M}_1)}$$

- ▶ 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

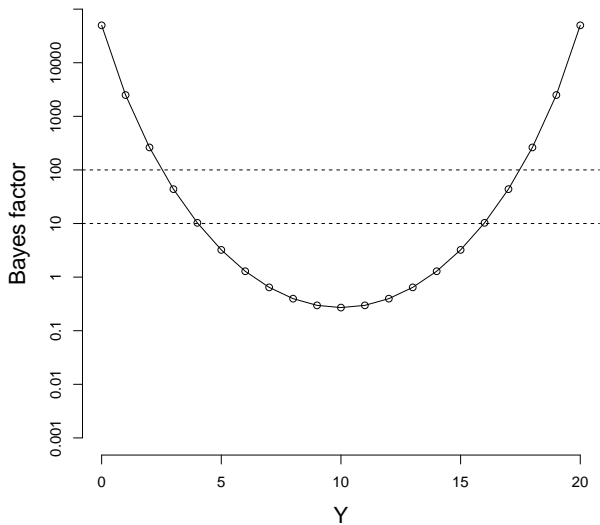
$$\mathcal{M}_1 : \theta = 0.5 \quad \text{and} \quad \mathcal{M}_2 : \theta \neq 0.5$$

- ▶ $p(Y|\mathcal{M}_1)$ is just the binomial density with $\theta = 0.5$
- ▶ \mathcal{M}_2 involves an unknown parameter θ
- ▶ This requires a prior, say $\theta \sim \text{Beta}(a, b)$, and integration

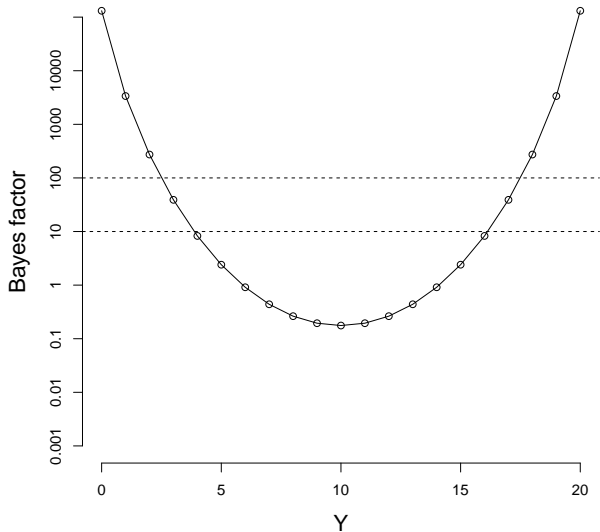
$$p(Y|\mathcal{M}_2) = \int p(Y, \theta) d\theta = \int p(Y|\theta) p(\theta) d\theta$$

- ▶ See “BF Beta-binomial” in the online derivations

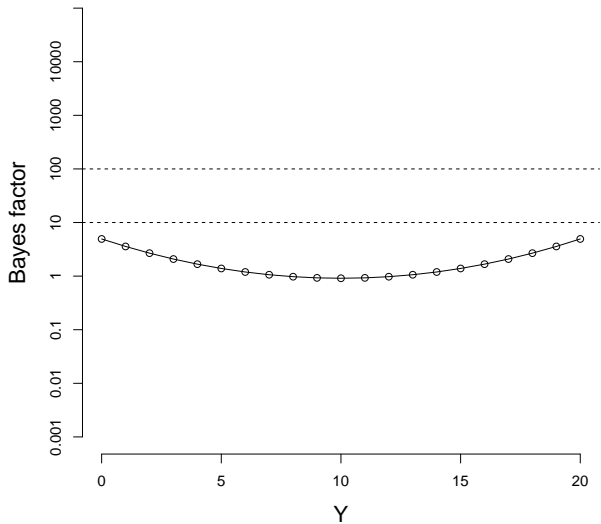
BF by Y with $n = 20$ and prior $a = 1$ and $b = 1$



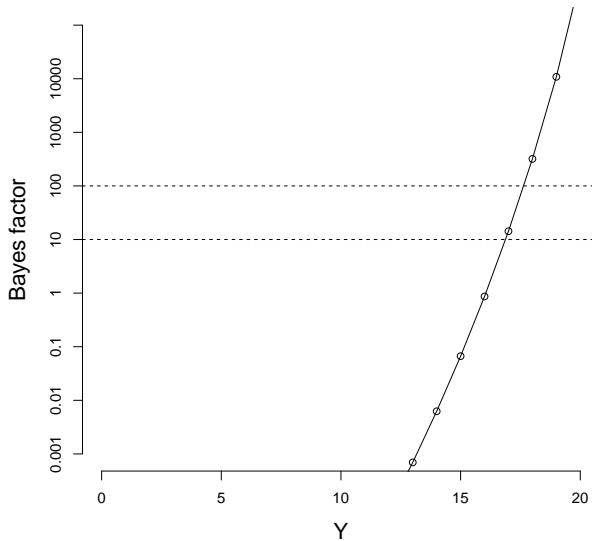
BF by Y with $n = 20$ and prior $a = 0.5$ and $b = 0.5$



BF by Y with $n = 20$ and prior $a = 50$ and $b = 50$



BF by Y with $n = 20$ and prior $a = 50$ and $b = 1$



Problems with Bayes factors

- ▶ Often hard to compute the required integrals which is only feasible for simple models
- ▶ Requires proper priors
- ▶ Can be very sensitive to priors (Lindley's paradox)
- ▶ In most cases, I prefer computing posterior intervals from the full model and testing by comparing these to the null

Computing Bayes factors using MCMC

- ▶ If models can be written as nested, then MCMC can be used to approximate model probabilities
- ▶ For example, say

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

- ▶ Both model can be written as

$$E(Y) = \beta_0 + \beta_1 X + \gamma \beta_2 X^2$$

where $\gamma \in \{0, 1\}$ indicates the model

- ▶ The prior on models becomes $\gamma \sim \text{Bernoulli}(0.5)$
- ▶ Then $\text{Prob}(\gamma = 1 | \mathbf{Y}) = \text{Prob}(\mathcal{M}_2 | \mathbf{Y})$ can be approximated using MCMC

Stochastic search variable selection (SSVS)

- ▶ This is the Bayesian analog of forward/backward/stepwise variable selection
- ▶ We place a prior on all 2^p models using p variable inclusion indicators γ_j
- ▶ MCMC returns the approximate posterior probability of each model
- ▶ With large p all models will have low probability and so this requires long MCMC runs
- ▶ As with Bayesian factors, SSVS can be sensitive to priors

Model averaging

- ▶ Let's go back to the linear regression example

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

- ▶ 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

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
- ▶ This is implemented in the R package `BMA`

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 predictions than competing models
- ▶ Advantages: Simple, intuitive, and broadly applicable
- ▶ Disadvantages: Slow because it requires several model fits and it is hard to say a difference is statistically significant

K-fold Cross validation

- 0 Split the data into K equally-sized groups
- 1 Set aside group k as test set and fit the model to the remaining $K - 1$ groups
- 2 Make predictions for the test set k based on the model fit to the training data
- 3 Repeat steps 1 and 2 for $k = 1, \dots, K$ giving a predicted value \hat{Y}_i for all n observations
- 4 Measure prediction accuracy, e.g.,

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

Variants

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

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

Information criteria

- ▶ Several information criteria have been proposed that do not require fitting the model several times
- ▶ Many are functions of the **deviance**, i.e., twice the negative log likelihood

$$D(\mathbf{Y}|\theta) = -2 \log[f(\mathbf{Y}|\theta)]$$

- ▶ Ideally, models will have small deviance
- ▶ However, if a model is too complex it will have small deviance but will be unstable (over-fitting)
- ▶ The Akaike information criteria has a complexity penalty

$$AIC = D(\mathbf{Y}|\hat{\theta}) + 2p$$

where $\hat{\theta}$ is the MLE

- ▶ Model with smaller *AIC* are preferred

Bayesian information criteria (BIC)

- ▶ The Bayesian information criteria is similar

$$BIC = D(\mathbf{Y}|\hat{\theta}) + \log(n)p$$

- ▶ This is motivated as an approximation to the log Bayes factor of the model compared to the null model
- ▶ However, this is only an asymptotic (large n) approximation
- ▶ With large n the prior is irrelevant, and so this is not satisfying to a subjective Bayesian

Deviance information criteria (DIC)

- ▶ *DIC* is a popular Bayesian analog of *AIC* or *BIC*
- ▶ Unlike *CV*, *DIC* requires only one model fit
- ▶ 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)

- ▶ Let $\bar{D} = E[D(Y|\theta)|\mathbf{Y}]$ be the posterior mean of the deviance
- ▶ Denote $\hat{\theta}$ as the posterior mean of θ
- ▶ The effective number of parameters is

$$p_D = \bar{D} - D(\mathbf{Y}|\hat{\theta})$$

- ▶ DIC can be written like *AIC*,

$$DIC = \bar{D} + p_D = D(\mathbf{Y}|\hat{\theta}) + 2p_D$$

- ▶ 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 smallest *DIC*

DIC

- ▶ 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, $p_D \ll p$ if there are strong priors
- ▶ For example, how many free degrees of freedom do we have with $\theta \sim \text{Beta}(1, 1)$ versus $\theta \sim \text{Beta}(1000, 1000)$?
- ▶ In some cases p_D has a nice closed form
- ▶ A few examples are worked out in “DIC” on the online derivations

DIC

- ▶ 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
- ▶ *DIC* can only be used to compare models with the same likelihood

Watanabe-Akaike information criteria (WAIC)

- ▶ *WAIC* is an alternative to *DIC*
- ▶ It is motivated as an approximation to leave-one-out CV
- ▶ In the end *WAIC* has model-fit and model-complexity components
- ▶ It is used the same as *DIC* with smaller *WAIC* begin preferred
- ▶ In practice the two often give similar results, but *WAIC* is arguably more theoretically justified

Watanabe-Akaike information criteria (WAIC)

- ▶ *WAIC* is written in terms of the posterior of the likelihood rather than parameters
- ▶ Let m_i be the posterior mean of

$$f(Y_i|\theta)$$

and v_i be the posterior variance of

$$\log[f(Y_i|\theta)]$$

- ▶ The effective model size is $p_W = \sum_{i=1}^n v_i$
- ▶ The criteria is

$$WAIC = -2 \sum_{i=1}^n \log(m_i) + 2p_W$$

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; added variable plots; etc.
- ▶ 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 then predict $Y_{new} \sim f(y|\hat{\theta})$
- ▶ This suppresses uncertainty in θ
- ▶ We would like to propagate this uncertainty through to the predictions

Posterior predictive distributions (PPD)

- ▶ We really want the PPD

$$f(Y_{new}|\mathbf{Y}) = \int f(Y_{new}, \theta|\mathbf{y})d\theta = \int f(Y_{new}|\theta)f(\theta|\mathbf{y})d\theta$$

- ▶ 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 datasets from the PPD with the identical design (same n , same \mathbf{X}) as the original data set

- ▶ We then define a statistic describing the dataset, e.g.,

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

- ▶ Denote the statistic for the original data set as d_0 and the statistic from simulated data set number s as d_s
- ▶ If the model is correct, then d_0 should fall in the middle of the d_1, \dots, d_S

Posterior predictive checks

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

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

- ▶ If p is near zero or one the model doesn't fit
- ▶ This is repeated for several d to give a comprehensive evaluation of model fit