Lecture 11B
MTH6102: Bayesian Statistical Methods

Eftychia Solea

Queen Mary University of London

2023

Today's agenda

Today's lecture

- Bayesian model selection

Next week

Revision next week

- Past papers
- Extra problems for the exam

More than one model

- Let y be the observed data.
- Suppose that we have two candidate statistical models that might fit the data y, models M_1 and M_2.
- Here, we assume that one of these models generated the data y.
- Each model has a vector of parameters θ_k, $k = 1, 2$.
- **Model selection:** We are interested in testing which model M_1 or M_2 fits the data y better.

Examples of more than one model

- Data: $y = (y_1, \ldots, y_n)$ (continuous).

$$M_1 : \; y_i \sim N(0, \sigma^2), \; \theta_1 = (\sigma) \quad \text{vs} \quad M_2 : \; y_i \sim N(\mu, \sigma^2), \; \theta_2 = (\mu, \sigma)$$

- We are interested in deciding whether or not μ is 0.

Examples of more than one model

- Regression models: $y_i \sim N(\mu_i, \sigma^2), i = 1, \ldots, n$, where σ is known.

  $M_1: \; \mu_i = \beta_0, \; \theta_1 = (\beta_0, \sigma)$ vs $M_2: \; \mu_i = \beta_0 + \beta_1 x_{1i}, \; \theta_2 = (\beta_0, \beta_1, \sigma)$

- We are interested in deciding whether or not β_1 is 0.

Hypothesis tests: frequentist

- In the frequentist framework, we have a null and alternative hypothesis.

$$H_0 : \mu = 0 \quad H_1 : \mu \neq 0$$

- Test hypotheses using p-value: Probability of statistic at least as extreme as the observed value, if H_0 is true.

Posterior probabilities

- The Bayesian framework does not use p-values.
- Probability statements are based on the posterior distribution conditional on the model M_k, $k = 1, 2$

Notation for inference in one model

- Recall the Bayes' theorem

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{p(y)}$$

- Conditional on the model M_k, Bayes' theorem becomes

$$p(\theta_k \mid y, M_k) = \frac{p(\theta_k \mid M_k)\, p(y \mid \theta_k, M_k)}{p(y \mid M_k)}, \quad k = 1, 2$$

where

$$p(y \mid M_j) = \int p(\theta_j \mid M_j)\, p(y \mid \theta_j, M_j)\, d\theta_j, \quad j = 1, 2$$

This is the probability of the data given model M_j is true.

Bayes' Theorem

$$p(\theta|y) = \frac{p(\theta)\, p(y|\theta)}{p(y)}$$

Suppose now that the Model M_x is true, then the Bayes' Theorem

$$p(\theta_x|y, M_x) = \frac{p(\theta_x|M_x)\, p(y|\theta_x, M_x)}{p(y|M_x)}$$

where M_x has parameters θ_x.

Bayes' theorem among models

the likelihood of the observed data y given M_k is true

- The term $p(y \mid M_k)$ can be used in Bayes' theorem for looking probabilities of different models (hypotheses).
- Bayes' theorem for model M_k (hypothesis)

$$p(M_k \mid y) = \frac{p(M_k)\, p(y \mid M_k)}{p(y)}, \quad k = 1, 2$$

- $p(M_k \mid y)$ is the posterior probability that model M_k is correct given the data y.
- These probabilities add up to 1: $\sum_{k=1}^{2} p(M_k \mid y) = 1$
- This provides a Bayesian method for choosing between models M_1 and M_2

Posterior probability of each model

- Hypotheses: We are testing two models: model M_1 and model M_2
- Prior probability: The probability of each model M_k, $k = 1, 2$ prior to collecting the data. In this case, we have $p(M_1) + p(M_2) = 1$

$$p(M_1) \quad \text{and} \quad p(M_2).$$

- Data: the result of the experiment. In this case, y.
- Likelihood: The probability of the data given model M_j is true, $p(y \mid M_j)$. In this case,

$$p(y \mid M_1) \quad \text{and} \quad p(y \mid M_2),$$

where

$$p(y \mid M_j) = \int p(\theta_j \mid M_j)\, p(y \mid \theta_j, M_j)\, d\theta_j, \quad j = 1, 2$$

Posterior probability of each model

- Posterior probability: The probability of each model M_k given the data y. In this case,

$$p(M_1 \mid y) \quad \text{and} \quad p(M_2 \mid y).$$

- By Bayes' theorem,

$$p(M_k \mid y) = \frac{p(M_k)\, p(y \mid M_k)}{p(y)}, \quad k = 1, 2.$$

- The denominator is

$$p(\text{data}) = p(y) = \sum_{j=1}^{2} p(M_j)\, p(y \mid M_j).$$

Prior distribution for models

- We need to specify prior probabilities for each model, $p(M_j)$, $j = 1, 2$.

- We could choose a discrete uniform distribution

$$p(M_j) = \frac{1}{r},\ j = 1, 2.$$

- (But we do not have to choose this distribution)

$$\underbrace{p(M_1) = a \quad a \in [0,1] \quad p(M_2) = 1 - a}_{\text{any valid pmf}}$$

Two models

So, we have by Bayes' theorem,

$$p(M_k \mid y) = \frac{p(M_k)\, p(y \mid M_k)}{p(y)}, \quad k = 1, 2.$$

- Suppose we assume one of two models is correct, M_1 and M_2.
- We want to decide which model fits the data y well.
- We choose M_1 or not depending on whether its posterior odds are greater or less than its prior odds.

Odds

- The odds of event E versus event E^{C} are the ratio of their probabilities $P(E)/P(E^{\mathsf{C}})$.

- So the odds of E is

$$O(E) = \frac{P(E)}{P(E^{\mathsf{C}})}.$$

- Let $P(E) = p$ and $P(E^{\mathsf{C}}) = 1 - p$, then $O(E) = \frac{p}{1-p}$.

$O(E) = \frac{p}{1-p}$. We can solve for p.

$O(E)(1-p) = p \iff O(E) - pO(E) = p$

$O(E) = p(1 + O(E)) \iff p = \frac{O(E)}{1+O(E)}$

Odds:Examples

- For a fair coin the odds of H (heads) is $O(H) = 1$. We say the odds of heads are 1 to 1 or 50-50.

- For a standard die, the odds of rolling 4 are $\frac{1/6}{5/6} = 1/5$. We say that odds are 1 to 5 for rolling a 4.

Prior odds, posterior odds

$p(M_1) + p(M_2) = 1$

- We compute,

$$\frac{p(M_1 \mid y)}{p(M_2 \mid y)} = \frac{p(M_1)\, p(y \mid M_1)}{p(M_2)\, p(y \mid M_2)}$$

- Also

$$p(M_2) = 1 - p(M_1),$$
$$p(M_2 \mid y) = 1 - p(M_1 \mid y)$$

$p(M_1|y) + p(M_2|y) = 1$

$$p(M_1|y) = \frac{p(M_1)\,p(y|M_1)}{p(y)}$$

$$p(M_2|y) = \frac{p(M_2)\,p(y|M_2)}{p(y)}$$

Take the ratio

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{[p(M_1)\,p(y|M_1)]/\cancel{p(y)}}{[p(M_2)\,p(y|M_2)]/\cancel{p(y)}}$$

$$= \frac{p(M_1)\,p(y|M_1)}{p(M_2)\,p(y|M_2)}$$

Prior odds, posterior odds

- The prior odds of model M_1 vs model M_2:

$$\frac{p(M_1)}{p(M_2)} = \frac{p(M_1)}{1 - p(M_1)}$$

- The posterior odds of model M_1 vs model M_2:

$$\frac{p(M_1 \mid y)}{p(M_2 \mid y)} = \frac{p(M_1 \mid y)}{1 - p(M_1 \mid y)}$$

Bayes factors

- Using,

$$\underbrace{\frac{p(M_1 \mid y)}{p(M_2 \mid y)}}_{\text{posterior odds of Model } M_1} = \frac{p(M_1)\, p(y \mid M_1)}{\underbrace{p(M_2)}_{\text{prior odds of Model } M_1}\, p(y \mid M_2)}$$

we have

$$\text{posterior odds of Model } M_1 = \text{prior odds of Model } M_1 \times \frac{p(y \mid M_1)}{p(y \mid M_2)}$$

Bayes factors

- The factor

$$B_{12} = \frac{p(y \mid M_1)}{p(y \mid M_2)}$$

is called a Bayes factor.

- So the Bayes factor is the ratio of the likelihoods.
- We have:

Posterior odds of Model M_1 = prior odds of Model M_1 $\times$ **Bayes factor**

Bayes factors

- For a hypothesis H (e.g Model M_1) versus $H^{\complement}$ (e.g Model M_2), the Bayes factor is

$$B_{12} = \frac{p(y \mid H)}{p(y \mid H^{\complement})}$$

- We have:

$$\text{Posterior odds of } H = \text{prior odds of } H \times \textbf{Bayes factor}$$

Bayes factor formula

- The Bayes factor is

$$\begin{aligned} B_{12} &= \frac{p(y \mid M_1)}{p(y \mid M_2)} \\ &= \frac{\int p(\theta_1 \mid M_1)\, p(y \mid \theta_1, M_1)\, d\theta_1}{\int p(\theta_2 \mid M_2)\, p(y \mid \theta_2, M_2)\, d\theta_2} \end{aligned}$$

- $p(\theta_k \mid M_k)$ and $p(y \mid \theta_k, M_k)$ are the prior and likelihood for model M_k.

Bayes factors and strength of evidence

Posterior odds of Model M_1 = prior odds of Model M_1 $\times$ Bayes factor

- The Bayes factor tells us whether the data provides evidence for or against Model M_1 (hypothesis)
  - Bayes factor $B_{12} > 1$ suggests the posterior odds are greater than the prior odds. So the data provides evidence for model M_1 (hypothesis). Model M_1 is more probable.
  - Bayes factor $B_{12} < 1$ suggests the posterior odds are less than the prior odds. So the data provides evidence against model M_1 (hypothesis). Model M_2 is more probable.
  - If $B_{12} = 1$ then the prior and posterior odds are equal. So the data provides no evidence either way.

Bayes factors and strength of evidence

- Rules of thumb for the size of the Bayes factor have been suggested - no need to remember these.
- E.g.:

Range of B_{12}	Evidence
1 to $10^{-\frac{1}{2}}$	slight evidence against M_1
$10^{-\frac{1}{2}}$ to 10^{-1}	moderate evidence against M_1
10^{-1} to 10^{-2}	strong evidence against M_1
$< 10^{-2}$	decisive evidence against M_1

Example

$n=k$

- We flip a coin 5 times and observe $k = 5$ heads. We want to know if the coin is fair, or if it is biased towards heads. Let q be the probability of success.
- Let be two models M_1 and M_2

$$M_1 : k \sim \text{binomial}(5, 0.5), \quad M_2 : k \sim \text{binomial}(5, q).$$ $q > 0.5$

- We will use the Bayes factor to choose between Models M_1 and M_2.

Example: By definition the Bayes factor, B_{12}, of Model M_1 vs Model M_2 is

$$B_{12} = \frac{p(x \mid M_1)}{p(x \mid M_2)}, \text{ where}$$

$$p(x \mid M_1) = \int_0^1 p(q \mid M_1)\, p(x \mid q, M_1)\, dq$$

$$p(x \mid M_2) = \int_0^1 p(q \mid M_2)\, p(x \mid q, M_2)\, dq$$

- Model M_1, there are no parameters since $q = 0.5$. Therefore, there are no parameters to integrate over. So,

$$p(x \mid M_1) = \binom{n}{x} (0.5)^x (0.5)^{n-x} \quad (n = x)$$

$$= (0.5)^n (0.5)^0 = \underline{(0.5)^n} = (0.5)^5$$

- $p(x \mid q, M_2)$ is the probability of x successes under M_2 and given the probability of success is q

$$p(x \mid q, M_2) = \binom{n}{x} q^x (1-q)^{n-x} \underset{(n=x)}{=} \underline{q^n}$$

$p(q|M_2)$ is the prior of q under model M_2. We can assume $p(q|M_2) \sim \underbrace{beta(1,1)}_{\text{uniform}}$ $\alpha, \beta, \alpha > 0, \beta > 0$

$$p(x|M_2) = \int_0^1 \underline{p(q|M_2)}\, p(x|q, M_2)\, dq$$

$$= \int_0^1 1 \cdot q^n\, dq = \int_0^1 q^n\, dq = \frac{1}{n+1}$$

Thus, the Bayes factor is

$$B_{12} = \frac{p(x|M_1)}{p(x|M_2)} = \frac{(0.5)^n}{\frac{1}{n+1}} = (n+1)(0.5)^n$$

since $n = 5$,

$$B_{12} = 0.1875 < 1$$

we conclude that model M_2 is more probab

Sensitivity to prior

- Suppose that model M_1 has a single parameter $\theta_1 \in \mathbb{R}$.
- Prior distribution $\theta_1 \sim N(0, \sigma_0^2)$.
- $$p(y \mid M_1) = \int p(\theta_1 \mid M_1)\, p(y \mid \theta_1, M_1)\, d\theta_1$$
- In typical problems, the likelihood $p(y \mid \theta_1, M_1)$ approaches zero for θ_1 outside some range $(-A, A)$.
- For large enough σ_0

$$p(\theta_1 \mid M_1) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\theta_1^2/(2\sigma_0^2)} \approx \frac{1}{\sqrt{2\pi}\sigma_0} \text{ for } -A < \theta_1 < A$$

Sensitivity to prior

- Hence for large enough σ_0 (flat, uninformative prior for θ_1), the Bayes factor is

$$B_{12} \approx \frac{1}{\sqrt{2\pi}\sigma_0} \frac{\int p(y \mid \theta_1, M_1)\, d\theta_1}{\int p(\theta_2 \mid M_2)\, p(y \mid \theta_2, M_2)\, d\theta_2}$$

- So if e.g. we replace a very large σ_0 by $100\,\sigma_0$, then B_{12} is divided by 100.
- However, the posterior distribution within model M_1 will hardly change, as the posterior is approximately proportional to the likelihood for large σ_0.

Alternative approaches to model comparison

- Using Bayes factors and posterior probabilities of models can depend on the prior distributions, more so than inference within each model.
- There are alternatives for checking or comparing models which combine Bayesian and frequentist ideas.
- E.g. posterior predictive checks.
- We are not covering these.

More flexible model

- An alternative is: don't choose among models.
- Expand one model to make it flexible enough.
- Models with many parameters can be easier to deal with in the Bayesian framework:
  - conceptually, can go from joint posterior to marginal posterior distribution;
  - having slightly informative prior distributions helps if there is not enough data to estimate all parameters.