3.36pt

Bayesian Model Averaging

A Brief Introduction

Gerald P. Dwyer

Clemson University

April 2020

Outline

3.36pt

General problem

- We have several different theories to explain the same data
- They may be mutually exclusive or overlapping
- Data for some may be a subset of other hypotheses' data
- Often reveals itself as many possible regressors and we are not sure which, if any, are important
- We have seen that Bayesian analysis and Bayes factors are a coherent way to compare hypotheses

Bayesian comparison of hypotheses

- Bayesian analysis makes it easy to compare two non-nested hypotheses and examine what the data say about them
- Start from prior probabilities of the two mutually exclusive and exhaustive hypotheses, $p(H_1)$ and $p(H_2)$
- The posterior probability of H_1 conditional on y is $p(H_1 | y)$
- The posterior probability of H_1 is related to the prior probability by Bayes rule,

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

- Similarly, the posterior probability of H_2 is related to the prior probability by Bayes rule,

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

Bayesian comparison of hypotheses

- The prior odds ratio of H_1 relative to H_2 is just

$$\frac{p(H_1)}{p(H_2)}$$

- The posterior probability of H_1 is

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

- The posterior probability of H_2 is

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

- The posterior odds ratio of H_1 relative to H_2 is

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

Comparison of models in a Bayesian analysis

- The comparison of models is given by

$$\frac{p(\mathsf{H}_1 \,|\, y)}{p(\mathsf{H}_2 \,|\, y)} = \frac{p(y \,|\, \mathsf{H}_1)}{p(y \,|\, \mathsf{H}_2)} \frac{p(\mathsf{H}_1)}{p(\mathsf{H}_2)}$$

- The ratio

$$\frac{p(y \,|\, \mathsf{H}_1)}{p(y \,|\, \mathsf{H}_2)}$$

  is called the "Bayes Factor" and is the relative likelihood of the models

- Note: the relative likelihood is not evaluated only at the maximum

- An application would have to have a prior distribution for the underlying parameters under H_1 and H_2 and these parameters are integrated out numerically

A different problem

- Have many alternative ways to characterize data
- Models $M_r, r = 1, R$
- Not trying to decide which is correct
- Trying to decide best way to characterize data and weight to be given to the various models
- Models may be different regressions
  - and most commonly are

Bayesian Model Averaging

- Have prior probabilities of the different models

$$p\left(M_r\right)$$

- Can compute posterior probabilities of the different models in the obvious way

$$p\left(M_r|y\right) = \frac{p\left(y|M_r\right)p\left(M_r\right)}{\Pr\left(y\right)}$$

Bayesian Model Averaging in Regression I

- Suppose that each model is a regression of the form

$$y_d = \beta_r x_r + \varepsilon_r$$

where y_d is a vector with a common dependent variable in the regressions

β_r is a vector of parameters

x_r is the matrix of data for model r

y is a matrix including y_d and all variables in any x

and ε_r is a vector of errors for model r

- Suppose that each model includes an estimate of the parameter β_i for variable i in x
  - The estimate might be zero because the variable is not included in x_r
- Then the expected value of β_i across the models is

$$\mathsf{E}\,\beta_i = \sum_{r=1}^{R} p\left(\beta_{i,r}|y, M_r\right) p\left(M_r\right) \beta_i$$

- That really is all there is all there is to it conceptually

Practical Issues in Bayesian Model Averaging in Regression I

- Cannot compute the posterior probabilities or the posterior expected values exactly
  - Use Markov Chain Monte Carlo to sample distribution in usual setup, for example regression
  - Obviates the need to have an exact representation of the posterior distribution
- Here also need to compute the posterior probability of models
- Markov Chain Monte Carlo Composition

Practical Issues in Bayesian Model Averaging in Regression I

- The weight on the variables and regressions must be chosen carefully
- A default option would be a flat prior for all equations
- This has what generally are undesirable properties
  - There are 2^K possible combinations of variables, where K is the number of variables
  - Suppose have 35 variables
  - The prior probability for each model is $p(M_\gamma) = 2^{-K} = 2.91 \times 10^{-11}$
  - The implied prior expected value of the length of the regressions is

  $$\sum_{k=0}^{K} \binom{K}{k} k 2^{-K} = K/2$$

  - Each equation gets the same prior probability which means that the probability of model size puts more weight on the intermediate model sizes
  - If that is what is desired, that is fine but it may not be very plausible

Practical Issues in Bayesian Model Averaging in Regression II

▶ With this setup, there 34,359,738,368 possible equations and equally weighting the equations puts a probability of 2.91×10^{-11} on each of these possible models

▶ This mean that different sizes of models get very different prior probability

▶ There are only 35 models with 1 variable, so all equations with 1 variable are quite unlikely given the extremely low probability of any particular model

▶ There is only 1 model with 35 variables, so this is very unlikely also

▶ Equations with intermediate numbers of variables receive the highest probability because they can include various combinations of the variables

   ★ The joint prior probability of the models with 17 variables is $\binom{K}{17} 2^{-K} = 13.2$ percent
   ★ The joint prior probability of the models with 18 variables is the same
   ★ The joint prior probability of models with 5 or fewer variables is 1.12×10^{-5}

Practical Issues in Bayesian Model Averaging in Regression III

★ The joint prior probability of models with 31 to 35 variables also is small, 1.73×10^{-6}.

- Alternative: have a prior expected length of regressions and use the expected length to infer probabilities of models
- Put equal weight on each variable given the prior expected regression length
  - ▶ Prior probability of each variable determined by expected length combined with the number of variables
  - ▶ The binomial prior has a common and fixed inclusion probability for each variable, say θ
  - ▶ The prior probability of a model of size k_γ is

$$p\left(M_\gamma\right) = \theta^{k_\gamma}\left(1-\theta\right)^{K-k_\gamma}$$

  - ▶ The expected model size is

$$\mathrm{E}\, m = K\theta.$$

Practical Issues in Bayesian Model Averaging in Regression IV

- ▶ Therefore, we can simply elicit a prior for $\mathsf{E}\, m$ and then infer

$$\theta = \frac{\mathsf{E}\, m}{K}.$$

  - ★ With $K = 35$ and $\mathsf{E}\, m = 5$, $\theta = 1/7 \approx 0.14$
  - ★ With $K = 35$ and $\mathsf{E}\, m = 6$, $\theta = 6/35 \approx 0.17$
  - ★ With $K = 35$ and $\mathsf{E}\, m = 17$, $\theta = 17/35 \approx 0.49$
  - ★ With $K = 35$ and $\mathsf{E}\, m = 25$, $\theta = 25/35 \approx 0.71$

- Other prior required is for the distribution of coefficients including the variance-covariance matrix
- Assume a diffuse prior for constant term – any value on the real line is equally likely
- Assume a normal prior distribution for the slope coefficients
  - ▶ Assume zero mean for distribution
  - ▶ Variance-covariance matrix

Practical Issues in Bayesian Model Averaging in Regression V

- Interesting and commonly used solution proposed by Zellner: use the actual variance-covariance matrix of the data scaled by a parameter g for the prior
  - Parameter g determines how much weight put on data compared to prior
  - This might seem inconsistent with having a prior because using the data in the prior, but it is not a serious issue in this instance because the posterior conditions on the same variables

What determines output losses?

- Will shift to PowerPoint