Bayesian Basics

Econ 690

Purdue University

Outline

Why Bayes?

- If you are like most, prior to coming to graduate school yet after the completion of a course or two in econometrics / statistics, your interpretation of a confidence region was probably correct from a Bayesian point of view, but incorrect from a frequentist perspective.

- A decision maker, for example, somewhat knowledgable about statistics and probability, may want to know what are the odds that a parameter lies in one region versus another. A frequentist finds providing an answer to such a question rather difficult.

- Before seeing the data, a 95% confidence region will contain the true parameter 95% of the time in repeated sampling. However, for decision making, don't we want to use the data that we see to help render a decision? After seeing the data, the realized region either contains the parameter or it does not.

Why Bayes?

Consider the following scenario ...

As the editor of the prestigious and widely-read journal, Bayes Stuff, you are charged to render a decision on a paper submitted for potential publication. A quick skim of the paper leads you to identify two equally qualified referees, A and B.

Unable to decide which referee to choose, you flip a coin which serves to identify referee B as the winner (or, perhaps, loser). Nine months later, referee B submits his / her decision to you. (Examples should be as realistic as possible, after all).

Why Bayes?

The relevant question is: for the purpose of making a decision on the paper's suitability for publication, should you also consider what referee A *might have said* had the coin selected him / her instead?

According to the pure frequentist prescription, you should, but I think most would agree that the decision should be made on the information available. In practice, this seems to always be the case.

Why Bayes?

There are a variety of examples pointing out holes / deficiencies of the frequentist approach ($p-$values as a guide to model selection, valid null confidence regions, etc.).

Admittedly, there are similar kinds of examples calling into question Bayesian reasoning, particularly those based on improper priors.

For many of you, the primary value of the course will be the set of tools we cover in the second half. This includes the description of *Markov Chain Monte Carlo* methods, which have proven to be quite useful for the fitting of all kinds of different models.

Why Bayes?

As a result of MCMC, we see more and more "Bayesians of necessity" out there, who wear the Bayesian hat temporarily (and, perhaps, with hope that they will not be seen) simply because MCMC requires them to do so.

After this course, some of you may join this growing group in the profession, leaving me both happy and sad. Hopefully, however, you will come through the other side more informed than most.

Review of Basic Framework

- Quantities to become known under sampling are denoted by the T-dimensional vector y.
- The remaining unknown quantities are denoted by the k-dimensional vector $\theta \in \Theta \subseteq \mathcal{R}^k$.
- Consider the joint density of observables y and unobservables θ:

$$p(y, \theta)$$

Review of Basic Framework

- Standard manipulations show:

$$p(y, \theta) = p(\theta)p(y|\theta) = p(y)p(\theta|y),$$

where

  - $p(\theta)$ is the prior density
  - $p(\theta|y)$ is the posterior density
  - $p(y|\theta)$ is the likelihood function. [Viewed as a function of θ, we write this as $L(\theta)$].

Review of Basic Framework

We also note

$$\begin{aligned} p(y) &= \int_{\Theta} p(\theta)p(y|\theta)d\theta \\ &= \int_{\Theta} p(\theta)L(\theta)d\theta \end{aligned}$$

is the marginal density of the observed data, also known as the marginal likelihood)

Bayes Theorem

Bayes' theorem for densities follows immediately:

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta)L(\theta)}{p(y)} \\ &\propto p(\theta)L(\theta). \end{aligned}$$

- We focus on the posterior up to proportionality ($\propto$) on the right-hand side. Often, the kernel of the posterior will take on a familiar form, whence the normalizing constant of the posterior can be deduced.
- The shape of the posterior can be learned by plotting the right hand side of this expression when $k = 1$ or $k = 2$.
- In these cases, the normalizing constant $[p(y)]^{-1}$ can be obtained numerically (e.g., a trapezoidal rule or Simpson's rule).

Bayes Theorem

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

- Obtaining posterior moments or posterior quantiles of θ, however, requires the integrating constant, i.e., the marginal likelihood $p(y)$.
- In most situations, the required integration cannot be performed analytically.
- In simple examples, however, this integration can be carried out. Many of these cases arise when the likelihood belongs to the exponential family of densities and the prior is chosen to be conjugate.
- By "conjugacy," we mean that the functional forms of the prior and posterior are the same.

Bayesian Inference

- Bayesian inference refers to the updating of prior beliefs into posterior beliefs conditional on observed data.
- The "output" of a Bayesian approach is the joint posterior $p(\theta|y)$.
- From this distribution:
  1. Point estimates can be obtained (e.g., posterior means),
  2. Posterior intervals can be calculated (e.g., $\Pr(a \leq \theta_j \leq b|y) = p$)
  3. (Posterior) predictions can be formulated regarding an out-of sample outcome.
  4. Hypothesis tests can be implemented.

Bayesian Inference

- It is important to keep in mind that, in the Bayesian viewpoint, θ is a random quantity, whereas statistics of the data, like

$$\hat{\theta} = (X'X)^{-1}X'y,$$

  which do not depend on θ, are not random *ex post*.
- It is also important to incorporate proper conditioning in your notation. For example,
  1. $E(\theta)$ refers to the prior mean of θ, whereas
  2. $E(\theta|y)$ refers to the posterior mean of θ.

Nuisance Parameters

In most practical situations not all elements of θ are of direct interest. Let

$$\theta = [\theta_1' \ \theta_2']' \in \ \Theta_1 \times \Theta_2,$$

and suppose that θ_1 denotes the parameters of interest while θ_2 are nuisance parameters.
For example θ_1 may be the mean and θ_2 the variance of some sampling distribution.
While nuisance parameters can be troublesome for frequentists, the Bayesian approach handles them in a natural way:

$$p(\theta_1|y) = \int_{\Theta_2} p(\theta_1, \theta_2|y) \ d\theta_2, \ \ \theta_1 \in \Theta_1.$$

i.e., they are simply marginalized (integrated out) of the problem.

Point Estimation

- As stated before, the output of a Bayesian procedure is the joint posterior $p(\theta|y)$.
- Coupled with a loss function $C(\hat{\theta}, \theta)$, which describes the loss of using $\hat{\theta}$ when the "true" parameter is θ, one can determine a point estimate $\hat{\theta}$ which minimizes posterior expected loss.
- These loss functions are commonly chosen to be increasing in the sampling error $\hat{\theta} - \theta$. (Whether these are symmetric or asymmetric will depend on the problem at hand).

Point Estimation

- A variety of popular loss functions include:
  1. quadratic loss,
  2. linear (absolute) loss,
  3. all-or-noting loss,
- which produce the
  1. posterior mean
  2. posterior median and
  3. posterior mode
- as the resulting point estimates, respectively.

Point Estimation

- In practice, one cannot provide a plot of $p(\theta_j|y)$ for all parameters in the model.
- So, how should one report the results of a Bayesian procedure?
- An applied Bayesian econometrician typically reports tables to summarize the posterior output – much like a frequentist would – supplying $E(\theta_j|y)$ and $\text{Std}(\theta_j|y)$ as summary statistics.
- Additionally, quantities like $\Pr(\theta_j > 0|y)$ are sometimes reported, and are superficially similar to the frequentist $p - value$ (which alienates frequentists and Bayesians alike). However, there are
  1. No stars!!!!
  2. No t-statistics!!!!!
- in the Bayesian's tables.

Interval Estimation

- Given a particular region $\Theta^* \subseteq \Theta$, one can immediately determine the *posterior probability that θ is contained in Θ^**:

$$\Pr(\theta \in \Theta^* | y) = \int_{\Theta^*} p(\theta|y) d\theta.$$

- For example:
  1. The probability that test scores decline with increases in class size from the regression model:

$$Score_i = \beta_0 + \beta_1 ClassSize_i + \epsilon_i$$

  2. The probability that a time series is stationary:

$$y_i = \alpha_0 + \alpha_1 y_{t-1} + \epsilon_t.$$

  3. The probability that a production function exhibits increasing returns to scale:

$$\log y_i = \beta_0 + \beta_1 \log L_i + \beta_2 \log K_i + \epsilon_i.$$

Interval Estimation

- In other cases, the process is reversed. That is, a desired probability of content p is determined, and an interval of minimum length with posterior probability of content p is constructed. This is termed a highest posterior density (HPD) interval

Testing

- A potential advantage of the Bayesian approach is its unified treatment of testing hypotheses.
- Consider two competing models, denoted $\mathcal{M}_1$ and $\mathcal{M}_2$. (These can be nested or non-nested).
- Note that

$$p(\mathcal{M}_j|y) = \frac{p(y|\mathcal{M}_j)p(\mathcal{M}_j)}{p(y)},$$

where

1. $p(\mathcal{M}_j|y)$ is the posterior probability of model j.
2. $p(y|\mathcal{M}_j)$ is the marginal likelihood under model j.
3. $p(\mathcal{M}_j)$ is the prior probability of model j.

- ▸ Return to next slide

Testing

- It follows that Jump back to last equation

$$\frac{p(\mathcal{M}_1|y)}{p(\mathcal{M}_2|y)} = \frac{p(y|\mathcal{M}_1)}{p(y|\mathcal{M}_2)} \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}$$

- where
  1. $p(\mathcal{M}_1|y)/p(\mathcal{M}_2|y)$ is the posterior odds of Model 1 in favor of Model 2.
  2. $p(y|\mathcal{M}_1)/p(y|\mathcal{M}_2)$ is termed the Bayes factor
  3. $p(\mathcal{M}_1)/p(\mathcal{M}_2)$ is the prior odds in favor of Model 1.

Testing

- Although this approach to testing can be generally applied, its implementation often proves difficult in models of moderate complexity since:

$$p(y|\mathcal{M}_j) = \int_{\Theta_j} p(y|\mathcal{M}_j, \theta_j) p(\theta_j|\mathcal{M}_j) d\theta_j$$

  is often non-trivial to calculate.
- We will return to a variety of strategies for approximating the marginal likelihood, or providing a consistent estimate of this quantity, toward the second-half of the course.

Prediction

- Consider an out-of-sample value y_f, presumed to be generated by the sampling model $y_f|\theta$.
- The Bayesian posterior predictive density is obtained as follows:

$$p(y_f|y) = \int_\Theta p(y_f|\theta)p(\theta|y)d\theta,$$

  where it is assumed that y_f is independent of y given θ (as in random sampling).
- In many models, this integration can also be difficult to perform analytically. Later in the course, we will describe simulation-based procedures for calculating the posterior predictive.

Choice of Prior

- If you asked a "representative frequentist" what he/she doesn't like about Bayesians ...
- after a really, really long series of uttered indecencies and a variety of obscene gestures ...
- and even more waiting ...
- he/she would probably settle down to criticize the prior.

Choice of Prior

- The Bayesian would likely respond by saying:
  1. How much of classical econometrics is truly "objective"? (e.g., pre-test assumptions such as whether a covariate is stationary, etc.)
  2. In large samples, which is where you guys live anyway, the prior washes out (under reasonable conditions).
  3. A good Bayesian should spend some time performing a sensitivity analysis - describing how posterior results change as the prior changes.

Choice of Prior

1. A good Bayesian should also try and do a reasonably convincing job of picking "reasonable" priors. This can be done in a variety of different ways, including:
   - Carefully using results of previous work to formulate priors.
   - Spending some time on prior elicitation, often in terms of thinking predictively.
   - Using part of the data set (or a related data set) to obtain hyperparameters of the prior distributions.
   - Use "reference" priors which attempt to be minimally informative. (We won't spend too much time on these).
2. For purposes of estimation, when the data set is moderate or large, proper yet reasonably diffuse priors generally have little impact on the posterior results. However, for purposes of testing, the prior matters a great deal.

Choice of Likelihood

- Another issue the frequentist is likely to raise is the sensitivity of results to the choice of likelihood.
- Again, a "good" Bayesian should:
  1. Implement a variety of diagnostic checks to see if his/her assumptions are supported by, or at odds with, the data.
  2. Re-model if necessary. Toward the end of the course, we will introduce a variety of computationally tractable ways to allow for fat tails, skew and multimodality in the error distribution (if needed).

Modern Bayesian Econometrics

- Powerful computing and the adoption of (relatively) new simulation-based statistical tools have really increased the popularity of Bayesian methods.
- In many cases, such tools make it possible to fit models that are very difficult, if not virtually impossible, with traditional frequentist methods.
- The key players here are the Gibbs sampler and the Metropolis-Hastings algorithm, two examples of Markov Chain Monte Carlo (MCMC) methods. We will discuss these algorithms, and their uses in a variety of models of interest to economists, in (roughly) the second-half of the course.

Modern Bayesian Econometrics

- Before getting into these computational methods, it is important to first have an understanding of what Bayesian econometrics is all about. We will address each of the issues mentioned in these slides in more detail, including (among other issues)
  1. Prior-posterior analysis (general)
  2. Prior-posterior analysis (in the regression model)
  3. Point Estimation
  4. Interval Estimation
  5. Hypothesis Testing
  6. Prediction
  7. Large Sample Bayes (Briefly)