Hypothesis Testing

Econ 690

Purdue University

Outline

1. Basic Testing Framework

2. “Testing” with HPD intervals

3. Example

4. Savage Dickey Density Ratio

5. Bartlett’s Paradox

Basic Testing Framework

- An 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).
- For example, $\mathcal{M}_1$ can denote an (unrestricted) regression model with k explanatory variables, and $\mathcal{M}_2$ denotes $\mathcal{M}_1$ with one [or more] of the explanatory variables removed.
- Alternatively, $\mathcal{M}_1$ could denote a regression model with Gaussian (normal) errors while $\mathcal{M}_2$ denotes the model with Student-t errors.
- $\mathcal{M}_1$ could denote the probit model while $\mathcal{M}_2$ could denote the logit.

Basic Testing Framework

- 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.
4. $$p(y) = \sum_{j=1}^{J} p(y|M_j)\Pr(M_j).$$

Basic Testing Framework

- Consider two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$, and note from our previous derivation that

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

- 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.

Basic Testing Framework

$$K_{12} = \frac{p(\mathcal{M}_1|y)}{p(\mathcal{M}_2|y)}$$

- If the posterior odds ratio above equals unity, then we are indifferent between $\mathcal{M}_1$ and $\mathcal{M}_2$.
- Jeffreys (1961) recommends interpreting a Bayes factor exceeding (approximately) 10 as strong evidence in favor of $\mathcal{M}_1$ (and a Bayes factor smaller than .1 as strong evidence in favor of $\mathcal{M}_2$).
- Of course, the magnitude of the odds itself is interpretable, and one does not really need to *select* a model on the basis of this information.
- Actually, a more sensible thing to do is to average over the models using the corresponding probabilities as weights. We will return to such Bayesian model averaging later in the course.

Basic Testing Framework

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

- Later in the course, we will provide a variety of numerical, simulation-based approaches for approximating marginal likelihoods (and thus Bayes factors). In this lecture we will also describe an approach for doing this when the models involve a zero subvector hypothesis.

Intuitive Testing

- Consider the regression model

  (i.e., a regression equation for a production function).
- The primary parameter of interest is the return to scale parameter
- In previous slides on the linear regression model, we showed that under the prior

  $$p(\beta, \sigma^2) \propto \sigma^{-2}$$

  the marginal posterior for β is of the form:

  with X defined in the obvious way.

Intuitive Testing

- The posterior distribution of θ can be obtained by first defining the selector matrix (here a vector) R as follows:

- Using standard properties of the multivariate Student-t distribution, it follows that

- Thus, via a simple change of variable, we have derived the posterior distribution for the return to scale parameter θ.

Intuitive Testing

- Given this posterior, one can immediately calculate probabilities of interest, such as the posterior probability that the production function exhibits increasing returns to scale:

- with T_ν denoting the standardized Student-t cdf with ν degrees of freedom.

Intuitive Testing

- Alternatively, we can calculate a $1-\alpha$ HPD interval for θ.
- Since the marginal posterior for θ is symmetric around the mean / median / mode $R\hat{\beta}$, it follows that

  is a $1-\alpha$ HPD interval for θ.
- If this interval does not include 1 (with a reasonable choice of α), then there is little evidence supporting constant returns to scale.
- The above offers a reasonable way of investigating the credibility of various hypotheses.

Marginal Likelihoods in a Regression Example

- Consider, again, our regression example:

$$y_i = \beta_0 + \beta_1 Ed_i + \epsilon_i, \quad \epsilon_i | Ed \overset{iid}{\sim} N(0, \sigma^2).$$

- We seek to illustrate more formally how model selection and comparison can be carried out. For this regression model, we employ priors of the form

$$\begin{aligned} \beta|\sigma^2 &\sim N(\mu, \sigma^2 V_\beta) \\ \sigma^2 &\sim IG\left[\frac{\nu}{2}, 2(\nu\lambda)^{-1}\right] \end{aligned}$$

  with $\nu = 6, \lambda = [2/3](.2), \mu = 0$ and $V_\beta = 10I_2$.
- The prior for σ^2 has a mean and standard deviation equal to .2.

Marginal Likelihoods in a Regression Example

- In our previous lecture notes on the linear regression model, we showed that, with this prior, we obtain a marginal likelihood of the form:

$$y \sim t\left(X\mu, \lambda(I + XV_\beta X'), \nu\right).$$

- Given values for the *hyperparameters* μ, V_β, λ and ν, we can calculate the marginal likelihood $p(y)$ [evaluated at the realized y outcomes].

Marginal Likelihoods in a Regression Example

- In our regression analysis, a question of interest is: do returns to education exist?
- To investigate this question, we let $\mathcal{M}_1$ denote the unrestricted regression model, and $\mathcal{M}_2$ denote the restricted model, dropping education from the right hand side. [Thus, X is simply a column vector of ones under $\mathcal{M}_2$.]
- We then calculate $p(y)$ with the given hyperparameter values, as shown on the last slide, under two different definitions of X.

Marginal Likelihoods in a Regression Example

- When doing these calculations (on the log scale!), we obtain:
  1. $\log p(y|\mathcal{M}_1) = -936.7$
  2. $\log p(y|\mathcal{M}_2) = -1,021.4.$
- Since

$$\log K_{12} = \log[p(y|\mathcal{M}_1)] - \log[p(y|\mathcal{M}_2)] \approx 84.7,$$

$$K_{12} = \exp(84.7) \approx 6.09 \times 10^{36} \ (!)$$

- Thus, our data show strong evidence in favor of the unrestricted model, i.e., a model including education in the regression equation.
- This was obvious, perhaps, since the posterior mean of the return to education was .091 and the posterior standard deviation was .0066.

Savage Dickey Density Ratio

- Consider the likelihood $L(\theta)$, $\theta = [\theta_1' \ \theta_2']'$, and the hypotheses

- Let $p(\theta_1|\mathcal{M}_1)$ be the prior for parameters under $\mathcal{M}_1$
- and similarly, let $p(\theta_1, \theta_2|\mathcal{M}_2)$ be the prior under $\mathcal{M}_2$.
- Assume, additionally, that the priors have the following structure, generated from $g(\theta_1, \theta_2) = g(\theta_1|\theta_2)g(\theta_2)$:
  1.
  2.
- In other words, the prior we use for the restricted model $\mathcal{M}_1$ is the same prior that we would use under $\mathcal{M}_2$ given the restriction. (If prior independence is assumed in g, then we simply require the same priors for parameters common to both models).

Savage Dickey Density Ratio

Show, under these assumptions [Verdinelli and Wasserman (1995)] that the Bayes factor for $\mathcal{M}_1$ versus $\mathcal{M}_2$ can be written as the ratio (known as the Savage-Dickey density ratio) :

-

where

-

In other words, the test can be carried out by calculating the marginal posterior and marginal prior ordinates at $\theta_2 = 0$ under the unrestricted $\mathcal{M}_2$.

Savage Dickey Density Ratio

- If the posterior ordinate at zero is much higher than the prior ordinate at zero, (leading to $K_{12} > 1$), this implies that the data have moved us in the direction of the restriction. As such, it is sensible that we would support the restricted model.

- Note that everything here is calculated under the unrestricted $\mathcal{M}_2$, not unlike what we do when implementing a Wald test.

- Potentially this offers a much easier way to calculate (nested) hypothesis tests. Here we do not need to calculate the marginal likelihood (which is typically not possible analytically).

Savage Dickey Density Ratio

Integrating the joint posterior with respect to θ_1 yields

-

Savage Dickey Density Ratio

Dividing this expression by $p(\theta_2|\mathcal{M}_2)$ and evaluating both at $\theta_2 = 0$ implies

-

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- In the case of a diffuse prior $p(\theta) \propto c$, for some constant c, we have

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

- Thus, the posterior mode and the maximum likelihood estimate will agree.

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- For purposes of obtaining the posterior distribution, specification of an improper prior of the form

  $$p(\theta) \propto c$$

  still can lead to a proper posterior since

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

  i.e., the arbitrary constant c cancels in the ratio.

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- However, for purposes of testing, improper priors should generally be avoided.
- (They can, however, be employed for nuisance parameters which are common to both models under consideration).
- To see why improper priors should not be employed (when testing), consider $\mathcal{M}_1$ and $\mathcal{M}_2$ with priors $p(\theta_1) \propto c_1$ and $p(\theta_2) \propto c_2$.
- Then,

$$\frac{p(y|\mathcal{M}_1)}{p(y|\mathcal{M}_2)} = \frac{\int_{\Theta_1} p(y|\theta_1)p(\theta_1)d\theta_1}{\int_{\Theta_2} p(y|\theta_2)p(\theta_2)d\theta_2} = \frac{c_1}{c_2}\frac{\int_{\Theta_1} p(y|\theta_1)d\theta_1}{\int_{\Theta_2} p(y|\theta_2)d\theta_2}$$

involves a ratio of arbitrary constants c_1/c_2.

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- OK. So, improper priors should be avoided when testing.
- To get around this issue, can we instead employ a proper prior (i.e., one that integrates to unity) and let the prior variances be huge?
- Wouldn't our associated hypothesis test then be (nearly) free of prior information?
- And give us a similar result to a classical test (since the posterior is "nearly" proportional to the likelihood)?
- The following exercise illustrates, perhaps surprisingly, that this is not the case, and that testing results are quite sensitive to the prior.

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- Suppose

  $$Y_t|\theta \overset{iid}{\sim} N(0,1)$$

  under hypothesis H_1 and

- $$Y_t|\theta \overset{iid}{\sim} N(\theta,1)$$

  under hypothesis H_2.

- In this example, we restrict the variance parameter to unity to fix ideas, and focus attention on scalar testing related to θ.
- Assume the prior

  $$\theta|H_2 \sim N(0,v), \;\; v > 0.$$

- Find the Bayes factor K_{21} (i.e., the odds in favor of allowing a non-zero mean relative to imposing a zero mean) and discuss its behavior for large v.

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- First, note that under H_1 there are no unknown parameters, and so the likelihood and marginal likelihood functions are the same.
- (This is like a dogmatic prior that imposes the mean to be zero and the variance to be unity with probability one). Thus, when integrating the likelihood over the "prior" we simply evaluate the likelihood at these values.
- Thus,

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- To evaluate the marginal likelihood under H_2, we must calculate:

- As before, we must complete the square on θ and then integrate it out.

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- Note that

$$
\begin{aligned}
\sum_t (y_t - \theta)^2 + v^{-1}\theta^2 &= \sum_t y_t^2 - 2\theta T\overline{y} + \theta^2(T + v^{-1}) \\
&= [T + v^{-1}]\left(\theta^2 - 2\theta\frac{T\overline{y}}{T + v^{-1}} + \frac{\sum_t y_t^2}{T + v^{-1}}\right) \\
&= [T + v^{-1}]\left[\left(\theta - \frac{T\overline{y}}{T + v^{-1}}\right)^2 - \left[\frac{T\overline{y}}{T + v^{-1}}\right]^2\right] + \\
&\quad\ [T + v^{-1}]\left(\frac{\sum_t y_t^2}{T + v^{-1}}\right).
\end{aligned}
$$

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- Plugging this back into our expression for the marginal likelihood, we obtain:

- The last line is *almost* the integral of a normal density for θ, except we need an integrating constant equal to $[T + v^{-1}]^{1/2}$

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- So, multiplying and dividing by $[T + v^{-1}]^{1/2}$, we obtain a marginal likelihood under H_2 equal to

- Write $[T + v^{-1}]^{-1/2}$ as

$$\begin{aligned} [T + v^{-1}]^{-1/2} &= (T^{-1}v)^{1/2}\left[T^{-1}v\left(T + v^{-1}\right)\right]^{-1/2} \\ &= T^{-1/2}v^{1/2}(v + T^{-1})^{-1/2}. \end{aligned}$$

- Thus, the term outside the exponential kernel of our marginal likelihood can be written as

$$(2\pi)^{-T/2}(v + T^{-1})^{-1/2}T^{-1/2}.$$

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- Recall that

$$p(y|H_1) = (2\pi)^{-T/2} \exp\left(-\frac{1}{2}\sum_t y_t^2\right).$$

- and we have shown

$$p(y|H_2) = (2\pi)^{-T/2}(v+T^{-1})^{-1/2}T^{-1/2} \exp\left(-\frac{1}{2}\left[\sum_t y_t^2 - \frac{T^2\overline{y}^2}{T+v^{-1}}\right]\right)$$

Thus,

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

$$K_{21} = (v + T^{-1})^{-1/2} T^{-1/2} \exp\left(\left[\frac{vT}{1+vT}\right]\frac{T\overline{y}^2}{2}\right)$$

- Note that, as $v \to \infty$ (keeping T fixed but moderately large):
- The exponential kernel approaches

  and similarly
-

Bayes Factors under (Nearly) Flat Priors: Bartlett's Paradox

- The results of this exercise clearly show that results of the test will depend *heavily* on the value of v chosen.
- In particular, as v grows, we tend to prefer the restricted model H_1. ? This rather counter-intuitive result is known as Bartlett's paradox.
- The lesson here is that priors will matter considerably for purposes of testing, while for estimation, choice of prior is decidedly less important.