Introduction to Bayesian Econometrics I

Prof. Jeremy M. Piger

Department of Economics

University of Oregon

Last Revised: March 15, 2019

1 Preliminaries

- At the core of Bayesian methods is probability. We will use the following notation to denote probability density functions (pdf):
  - $p(.)$ is a pdf for a generic random variable. That is, we will interpret $p(y)$ as the pdf for the random variable y, $p(x)$ as the pdf for the random variable x, $p(x,y)$ as the joint pdf for the random variables x and y, and $p(x|y)$ as the conditional pdf for x given y.
  - If the random variable is discrete we will write the probability mass function (pmf) as $\Pr(\cdot)$.
  - We won't distinguish between random variables and realizations in notation. Which is which should be clear from the context and related discussion.
- Before going further, you should read the document "Review of Important Probability Density Functions."

2 Bayes' Rule and Bayesian Inference

- The **law of total probability** says that:

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

- Of course, we can reverse the right hand side of the above equation:

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

- **Bayes' Rule** follows immediately from these two equations:

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

- Suppose that x has taken on a value, but we do not know what it is. In other words, x is an **unknown quantity**. We would like to know more about the likely values for this unknown quantity. Suppose further that y has taken on a value, and we know what this value is. In other words, y is a **known quantity**. Bayes' Rule gives us a formula to express what we know about x, given what we know about y, as a conditional pdf. This pdf will allow us to make probability statements about the quantity x. For example, we could make a statement such as: "Given y, there is a 50% probability that $x > 2$."

- Bayesian econometrics applies this framework to econometric models. In particular, suppose we have a model describing a random variable y in terms of a k-dimensional set of parameters called $\theta \in \Re_k$. Suppose that θ is a random variable. To help fix ideas, suppose that θ is realized, which then plays a role in generating a value for y in accordance with the model. Finally, suppose that we can observe y via sampling. We collect a T-dimensional vector of observations on y, collected in the vector Y. On the other hand, we do not observe the value of θ. Bayes' Rule says that we can summarize what we know about θ (the parameters), conditional on having seen Y (the data), via the following equation:

$$p(\theta|Y) = \frac{p(Y|\theta)\,p(\theta)}{p(Y)} \tag{1}$$

- In this equation:

  - $p(\theta|Y)$ is the pdf for θ, conditional on the realized sample of observations Y. It summarizes what we know about θ after having seen the sample of observations on y. It is referred to as the "posterior" density.

- $p(Y|\theta)$ is the pdf for Y, conditional on a particular value for the parameters θ. It is functionally equivalent to the likelihood function for the model.

- $p(\theta)$ is the marginal pdf for θ. It summarizes what we know about θ without (or before) having seen the sample of observations on y. This is referred to as the "prior" density.

- $p(Y)$ is equivalent to the marginal pdf for Y, where we have integrated the joint pdf for Y and θ over θ:

  $$\begin{aligned} p(Y) &= \int_{\theta} p(Y, \theta)\, d\theta && (2) \\ p(Y) &= \int_{\theta} p(Y|\theta)\, p(\theta)\, d\theta && (3) \end{aligned}$$

  For this reason, $p(Y)$ is often referred to as the **marginal likelihood**. Note that it is equivalent to the expected value of the likelihood function, where the expectation is taken with respect to the prior for θ. As such, you will sometimes see the marginal likelihood referred to as the **average likelihood**. In some cases, most prominently when our only interest is in estimating θ, the marginal likelihood isn't of much interest. In other cases, most prominently when testing models, the marginal likelihood is of great interest. Much more on the marginal likelihood later.

- Bayes' Rule gives us the mechanism through which we conduct **Bayesian Inference** about the parameters of an econometric model.

- **Bayesian Inference:** the updating of prior beliefs into posterior beliefs conditional on observed data. The result of Bayesian Inference is the posterior density, $p(\theta|Y)$.

- Because $p(Y)$ is a constant, (1) implies:

$$p\left(\theta|Y\right) \propto p\left(Y|\theta\right) p\left(\theta\right)$$

This equation makes clear the how the prior density function is updated into the posterior density function. In particular, the prior density is updated by coming into contact with the data through the likelihood function.

- A **conjugate prior** refers to a prior density that, once it is interacted with the likelihood function, yields a posterior density of the same functional form as the prior.

- In some cases we may not be interested in all the parameters of the model. For example, we may be interested in the j^{th} element of θ, denoted θ_j. In this case, we can construct a marginal posterior distribution:

$$p\left(\theta_j|Y\right) = \int_{\theta_{\neq j}} p\left(\theta|Y\right) d\theta_{\neq j}$$

3 Example: Bernoulli Trials

- We begin with a simple example, which was the subject of the first published Bayesian analysis, by Thomas Bayes in 1763. Suppose we have a random variable, denoted y that is binary (either 0 or 1) and is assumed to have come from a Bernoulli distribution with parameter θ, where $0 < \theta < 1$. The probability mass function for this random variable is then:

$$\Pr\left(y\right) = \theta^y \left(1-\theta\right)^{1-y} \quad ; \quad y \in \{0,1\},$$

Now, suppose we don't observe θ and wish to estimate it based on a sample of T i.i.d. realizations from this Bernoulli distribution, denoted $Y = \left(y_1, y_2, \cdots, y_T\right)'$. The joint pmf for Y is then:

$$\Pr\left(Y|\theta\right) = \theta^{s}\left(1-\theta\right)^{T-s},$$

where $s = \sum_{i=1}^{T} y_i$

- To conduct Bayesian inference we require a prior density for θ. Here we will use the Beta distribution as the functional form for our prior density, so that $\theta \sim \text{Beta}\left(\alpha_1, \alpha_2\right)$, and:

$$p\left(\theta\right) = \frac{1}{B\left(\alpha_1, \alpha_2\right)}\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1} \text{ for } 0 < \theta < 1 \text{ and } \alpha_1, \alpha_2 > 0,$$

- This functional form is nice, because it explicitly incorporates the constraint that $0 < \theta < 1$.

- Once this functional form has been decided upon, the specification of the prior boils down to the choice of α_1 and α_2. Remember, that these values are chosen by you! They represent your priors.

- Before continuing, a quick note on Bayesian terminology. In a Bayesian analysis, the unknown parameters of the model that we wish to estimate are called **parameters**. The known parameters that are used in the specification of the prior distribution (above α_1 and α_2) are called **hyperparameters**.

- Suppose we choose values for the hyper parameters. The next step is to apply Bayes' rule. Recall, if we ignore the marginal likelihood, which does not depend on θ, then Bayes' rule implies:

$$p\left(\theta|Y\right) \propto p\left(Y|\theta\right)p\left(\theta\right)$$

For this example, we then have:

$$p\left(\theta|Y\right) \propto \left(\theta^{s}\left(1-\theta\right)^{T-s}\right)\left(\frac{1}{B\left(\alpha_1,\alpha_2\right)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}\right)$$

Further, since the Beta function doesn't depend on θ, we have:

$$p\left(\theta|Y\right) \propto \left(\theta^{s}\left(1-\theta\right)^{T-s}\right)\left(\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}\right)$$

Combining terms we have:

$$p\left(\theta|Y\right) \propto \left(\theta^{\alpha 1+s-1}\left(1-\theta\right)^{\alpha_2+T-s-1}\right)$$

- The previous equation will be recognized as the **kernel** (defined in the "Review of Important Probability Density Functions" document) of a Beta distribution with parameters $\overline{\alpha_1}=\alpha_1+s$ and $\overline{\alpha_2}=\alpha_2+T-s$. Thus, the posterior distribution is a Beta distribution and the prior and posterior distribution have the same functional form. The beta distribution is the conjugate prior for the probability of success parameter in a Bernoulli trial.
- The full formula for the posterior probability distribution function is then:

$$p\left(\theta|Y\right) = \frac{1}{B\left(\overline{\alpha_1},\overline{\alpha_2}\right)}\theta^{\overline{\alpha_1}-1}(1-\theta)^{\overline{\alpha_2}-1} \text{ for } 0<\theta<1$$

4 Objects of Interest from the Posterior

- The posterior density function contains all the information that we know about θ. One obvious approach to display evidence from the posterior is to plot each of the marginal posterior densities graphically. In practice, most Bayesian researchers alternatively / also present numerical information about the posterior density. As the posterior density function is a pdf, it can be used to flexibly construct any number of probability statements and summary statistics relevant to inference. Here we will review some of the most popular of these:

4.1 Point Estimates

- Using the posterior, we can construct a single, “point”, estimate of a parameter, denoted $\widehat{\theta}_j$.

- The Bayesian approach to a point estimate is based on a **loss function**, denoted $W\left(\widehat{\theta}_j, \theta_j\right)$. This describes the loss of using $\widehat{\theta}_j$ when the true parameter is θ_j.

- With a loss function in hand, we can then determine a point estimate, denoted $\widehat{\theta}_j$, that minimizes posterior expected loss, where the expectation is taken with respect to θ_j:

$$
\begin{array}{c}
E\left[W\left(\widehat{\theta}_j, \theta_j\right)|Y\right] \\
\int_{\theta_j} W\left(\widehat{\theta}_j, \theta_j\right) p\left(\theta_j|Y\right) d\theta_j
\end{array}
$$

- The three most common Bayesian point estimators are the posterior mean, posterior median, and posterior mode:

$$
\begin{aligned}
\text{posterior mean: } \widehat{\theta}_j &= E\left[p\left(\theta_j|Y\right)\right] \\
\text{posterior median: } \widehat{\theta}_j &= \text{median of } \ p\left(\theta_j|Y\right) \\
\text{posterior mode: } \widehat{\theta}_j &= \text{mode of } \ p\left(\theta_j|Y\right)
\end{aligned}
$$

- The posterior mean is the optimal estimator when the loss function is a **weighted squared error** loss function:

$$\text{posterior mean: } W\left(\widehat{\theta}_j, \theta_j\right) = Q\left(\theta_j - \widehat{\theta}_j\right)^2$$

  where Q is a positive scalar.

- The posterior median is the optimal estimator when the loss function is an **absolute** loss function, also called a **symmetric linear** loss function:

$$\text{posterior median: } W\left(\widehat{\theta}_j, \theta_j\right) = \left|\theta_j - \widehat{\theta}_j\right|$$

- The posterior mode is the optimal estimator when the loss function is an **all or nothing** loss function:

$$\text{posterior mode: } W\left(\widehat{\theta}_j, \theta_j\right) = \begin{cases} w & \text{if} \left|\theta_j - \widehat{\theta}_j\right| > d \\ 0 & \text{if} \left|\theta_j - \widehat{\theta}_j\right| \leq d \end{cases}$$

  and the error threshold d is taken to zero.

4.2 Measures of Posterior Uncertainty

- Point estimation simply reports a single estimate for θ. We might in addition be interested in the uncertainty about θ present in the posterior density function.

- Such uncertainty measures are usually constructed for individual parameters, θ_j.

- One obvious summary statistic of uncertainty, which is commonly reported, is the standard deviation of the marginal posterior distribution function for θ_j.

- Another way to summarize posterior uncertainty is by constructing a range of likely values for θ_j. This is called **interval estimation**.

- One way to construct interval estimates is by reporting a **central posterior interval**, which is an interval of values for θ_j such that an equal, pre-specified, amount of probability lies above and below the interval. Formally, if we desire a $100(1-\alpha)\%$ interval for θ_j, we would report the region of values for θ_j such that there is $100(\alpha/2)\%$ of the posterior probability above and below the region. The endpoints of this interval are simply the $\alpha/2$ and $(1-\alpha/2)$ percentiles.

- A slightly different method of summarizing posterior uncertainty is to compute a **highest posterior density (HPD) region**. This is defined as the region of values that contains $100(1-\alpha)\%$ of the posterior probability and also has the characteristic that the density within the region is never lower than that outside. Such a region will be identical to the central posterior interval if the posterior distribution is unimodal and symmetric. However, an HPD region can be quite different from the central posterior interval in other cases. For example, for a bimodal posterior distribution, the HPD region need not be a single interval, but could be two disjoint intervals.

- To formally define the HPD region, begin by considering the probability that θ_j lies inside the region defined by Θ:

$$\Pr\left[\theta_j \in \Theta | Y\right] = \int_{\Theta} p\left(\theta_j | Y\right) d\theta_j = 1 - \alpha$$

  The region Θ is defined as a $\mathbf{100\left(1-\alpha\right)\%}$ **Bayesian Credible Region**.

- An HPD region reverse engineers a Bayesian credible region. A desired probability $1-\alpha$ is set, and then a Bayesian credible region with this probability is found. However, there will generally be more than one such interval. For example, if $p\left(\theta|Y\right)$ is a normal distribution, then there will be many 95% Bayesian credible regions.

- The HPD region selects from among the alternative Bayesian credible regions by selecting that region such that the marginal posterior distribution for θ_j is always lower outside the region than it is inside it.

- In general, the HPD region is more difficult to compute than the central posterior interval.

4.3 Reporting Bayesian Results

- When reporting estimates of θ from a Bayesian estimation, the usual approach is to present tables that have a point estimate and a measure of dispersion for each θ_j, much like one would do in the classical framework. For example, one might report the posterior mean/median, along with the standard deviation of the posterior distribution and/or a central interval or HPD region. In some cases, it may be feasible to provide plots of the marginal posterior distribution for each parameter. You will also sometimes see a plot of the prior distribution for each parameter as compared to the posterior, which helps us see how the data updated our beliefs. You will also see objects of interest from that particular paper.

5 Exercise: Bernoulli Trials

- Using the zipped collection of Matlab files titled “Bernoulli Trials.zip”, complete the following tasks:

  1. Read through the programs and make sure you understand what it is doing (the main program is “Bernoulli_Trials.m”. Run the program - make sure you understand the output.

  2. The data sample is defined on line 8 of the main program. Double the sample size, keeping the same proportion of 0’s and 1’s. How does this affect the results?

3. Play around with the hyperparameters of the prior density. How does this affect things?

6 Bayesian Prediction

- **Prediction**: We also use the posterior distribution, along with the model likelihood function, to draw inference regarding an out-of-sample prediction, or forecast, of y, denoted y^*. We do this by constructing the **posterior predictive density**:

$$p\left(y^*|Y\right) = \int_\theta p\left(y^*, \theta|\, Y\right) d\theta$$

- Using the law of total probability, we can write this as:

$$p\left(y^*|Y\right) = \int_\theta p\left(y^*|\theta, Y\right) p\left(\theta|Y\right) d\theta$$

- Using the posterior predictive density, we can form many of the same objects of interest discussed earlier for the posterior of θ. In particular, we can form a point prediction of y^* as the mean or median of $p\left(y^*|Y\right)$, or an HPD region for y^*.

- There are two notably desirable attributes of prediction using the posterior predictive density:

  - The posterior predictive density is a probability density function. As such, it immediately gives us a formal way of thinking not just about a point prediction, but also uncertainty about predictions. In the language of the forecasting literature, the posterior predictive density is a **density forecast**.

  - The posterior predictive density is not conditional on a particular value for θ. As such, inference taken from $p\left(y^*|Y\right)$ will incorporate uncertainty about parameters. For example, a measure of prediction uncertainty would incorporate uncertainty not just about future stochastic elements of the model, but also uncertainty about the value of the model's paramaters. This is unlike much of the non-Bayesian prediction / forecasting literature, where an estimate of θ is "plugged in" to create predictions.

7 Example: Bernoulli Trials

For the example of i.i.d. Bernoulli trials, we have

$$\begin{aligned} p\left(y^{*}|\theta, Y\right) &= p\left(y^{*}|\theta\right) \\ &= \theta^{y^{*}}\left(1-\theta\right)^{1-y^{*}} \end{aligned}$$

So, for the Bernoulli trials case considered earlier we have:

$$\begin{aligned} p\left(y^{*}|Y\right) &= \frac{1}{B\left(\overline{\alpha_1},\overline{\alpha_2}\right)}\int_0^1 \theta^{y^{*}}\left(1-\theta\right)^{1-y^{*}}\theta^{\overline{\alpha_1}-1}\left(1-\theta\right)^{\overline{\alpha_2}-1}d\theta \\ &= \frac{1}{B\left(\overline{\alpha_1},\overline{\alpha_2}\right)}\int_0^1 \theta^{y^{*}+\overline{\alpha_1}-1}\left(1-\theta\right)^{-y^{*}+\overline{\alpha_2}}d\theta \end{aligned}$$

The integrand in the second equation is the kernel of a Beta distribution with parameters $y^{*}+\overline{\alpha_1}$ and $-y^{*}+\overline{\alpha_2}+1$. Thus, we have:

$$\begin{aligned} p\left(y^{*}|Y\right) &= \frac{B\left(y^{*}+\overline{\alpha_1},-y^{*}+\overline{\alpha_2}+1\right)}{B\left(\overline{\alpha_1},\overline{\alpha_2}\right)} \\ &= \frac{B\left(y^{*}+\alpha_1+s,-y^{*}+\alpha_2+T-s+1\right)}{B\left(\alpha_1+s,\alpha_2+T-s\right)} \\ &= \frac{\Gamma\left(y^{*}+\alpha_1+s\right)\Gamma\left(-y^{*}+\alpha_2+T-s+1\right)}{\Gamma\left(\alpha_1+\alpha_2+T+1\right)}\frac{\Gamma\left(\alpha_1+\alpha_2+T\right)}{\Gamma\left(\alpha_1+s\right)\Gamma\left(\alpha_2+T-s\right)} \end{aligned}$$

We could use this formula to construct the posterior predictive probability that $y^{*}=0$ or $y^{*}=1$. For example, suppose that we have set our prior to be a uniform distribution for θ over $[0,1]$. In other words, $\alpha_1=1$ and $\alpha_2=1$. In this case, the arguments of the Gamma functions will be integers, and we can write:

$$p\left(y^{*}|Y\right)=\frac{\left(y^{*}+s\right)!\left(-y^{*}+T-s+1\right)!}{(2+T)!}\frac{(1+T)!}{(s)!\,(T-s)!}$$

Suppose we then are interested in the posterior predictive probability that $y^{*}=1$. This is given by:

$$\begin{aligned}p\left(y^{*}=1|Y\right) &= \frac{(1+s)!\,(T-s)!}{(2+T)!}\frac{(1+T)!}{(s)!\,(T-s)!}\\ p\left(y^{*}=1|Y\right) &= \frac{(1+s)}{(2+T)}\end{aligned}$$

By referring to the earlier equation for the posterior mean, we see that this is the equation for the posterior mean for θ, after plugging in $\alpha_1 = 1$ and $\alpha_2 = 1$. This is intuitive - the posterior predicted probability of the outcome of the next Bernoulli trial is simply the posterior expectation of the probability of success.

8 Example: Poisson Model

Count data are often modeled using a Poisson distribution with parameter λ, where $\lambda > 0$. Suppose we have a discrete random variable that records the count (number of times) something occurred. Thus, $y \in \{0,1,2,\cdots\}$. We assume that y comes from a Poisson distribution, so that the probability mass function for y is:

$$\Pr\left(y|\lambda\right)=\frac{\lambda^{y}e^{-\lambda}}{y!}\quad y\in\{0,1,2,\cdots\}$$

Now, suppose we don't observe λ and wish to estimate it based on a sample of N i.i.d. realizations from this Poisson distribution, denoted $Y=\left(y_1,y_2,\cdots,y_N\right)'$. The joint pmf for Y is then:

$$\Pr\left(Y|\lambda\right) = \frac{\lambda^{s} e^{-N\lambda}}{\prod\limits_{i=1}^{N} y_i!}$$

For the prior for λ we need a distribution that incorporates the constraint $\lambda > 0$. Suppose we choose a $\text{Gamma}\left(\alpha, \beta\right)$ for this purpose:

$$p\left(\lambda\right) = \frac{1}{\beta^{\alpha}} \frac{1}{\Gamma\left(\alpha\right)} \lambda^{\alpha-1} \exp\left[-\frac{\lambda}{\beta}\right]$$

9 Exercises: Poisson Model

1. Derive the formula for the posterior distribution of λ. Is a Gamma distribution the conjugate prior for this model?

2. Derive an equation for the posterior predictive density $p\left(y^{*}|Y\right)$, and show that this distribution is Negative Binomial.

3. Open the zipped collection of Matlab files titled "Poisson TrailBlazers.zip." This holds Matlab programs and data files to estimate a Poisson model for count data that is the number of points scored by the Portland Trail Blazers in games played during the 2014-2015 season through March 14, 2015. Data was obtained from "Basketball-Reference.com." Open the file "Poisson.m" and make sure you understand everything that is going on. The assumption to justify this estimation is that each of the number of points scored by the TrailBlazers are independent draws from a Poisson distribution with parameter $\lambda > 0$.

4. Think about how you would set the hyperparameters for a $\text{Gamma}\left(\alpha, \beta\right)$ prior for λ. Play around with different hyperparameters and see how it affects your results.

5. The program is also set up to use the posterior predictive distribution to compute

various predictions for a future game not yet played by the TrailBlazers. Inspect the code to understand what these predictions are. Implement some different predictions that you might be interested in.