ECO545A: Bayesian Econometrics

January 16, 2023

Chapter 1: Introduction

Econometrics is largely concerned with quantifying the relationship between one or more variables y, called the response or dependent variables, and one or more variable x, called regressors, independent variables or covariates. A general regression relationship can be expressed as $y = f(x) + \epsilon$.

y is continuous: Multiple linear regression models

y is discrete : binary models (e.g., probit, logit), discrete choice models, multinomial models, count data models (e.g., Poisson regression, negative binomial regression).

Recall: The linear regression model is given by the following equation,

$$y = X\beta + \epsilon,$$

where $y = (y_1, ..., y_n)'$ is a column vector of dimension $n \times 1$, X is a matrix of dimension $n \times k$ defined as,

$$X = \begin{pmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{nk} \end{pmatrix},$$

$\beta = (\beta_1, \cdots, \beta_k)'$ is a $k \times 1$ column vector of unknown parameters, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)'$ is $n \times 1$ vector of errors. The Ordinary least squares estimators are,

$$\hat{\beta} = (X'X)^{-1}(X'y), \qquad \text{and} \qquad \hat{\sigma}^2 = \hat{\epsilon}'\hat{\epsilon}/(n-k).$$

A couple of examples of regression are as follows:

1. y: quantities demanded of a set of goods
   x: income, prices and other characteristics of the goods
2. y: investment in a capital equipment
   x: measures of expected sales, cash flows and borrowing costs

In addition to covariates, there is a stochastic/random component (i.e., ϵ) to the regression. This makes y itself a random variable. The random variable y is characterized by a distribution.

When y is continuous, it is characterized by a *pdf* and denoted by $f(y|\theta, x)$. When y discrete, it characterized by a *pmf* and denoted by $f(y|\theta, x)$. It is customary to suppress the dependence on covariates. Hence, *pdf/pmf* will be denoted by $f(y|\theta)$.

Data for econometric modeling may be of different types:

1. *Cross section*: Observations on a number of subjects at the same point in time.
2. *Time Series*: Observations over a number of time periods.
3. *Panel Data*: Data over many subjects over a relatively short period of time.
4. *Multivariate Data*: Data over a fairly small number of subjects over long periods of time.

When modeling y as a function of x, the researcher may consider the covariates as fixed or as random variables. If the latter, their distribution may be either independent or dependent on the distribution of y.

Observational Data: In econometrics, the data are almost always observational, as opposed to data arising from controlled experiments, where subjects are randomly arranged to treatments.

Observational data may arise from surveys or as data collected by government for various reasons. In any case, the analysis of observational data requires special care, especially in the analysis of causal effects − the attempt to determine the effect of a covariate on a response variable where the covariate is a variable whose value can be set by an investigator, such as the effect of training program on income and employment or the effect of exercise on health. When data are collected by observing what people choose to do, rather than from controlled experiment, there is a possibility that people who choose are systematically different from people who do not. If so, attempting to generalizes the effect of training or exercise on people who do not freely choose those options may give misleading answers. This is the famous sample selection problem.

Chapter 2: Basic Concepts of Probability and Inference

A probability is a number $\in (0, 1)$ assigned to statements or events. Some examples of such statements are:

(1) A_1 : A coin tossed three times will show heads either two or three times.

(2) A_2 : A six sided dice rolled once shows an even number of spots.

Probability of an event A, denoted by $P(A)$ is assumed to satisfy the following axioms:

(1) $0 \leq P(A) \leq 1$

(2) $P(A) = 1$ if A represents a logical truth, i.e. a statement that must be true. Example: a coin when tossed comes up either tails or heads.

(3) If A and B are discrete disjoint events, then, $P(A \cup B) = P(A) + P(B)$.

(4) Let $P(A|B)$ denote the probability of A given that B is true. Then, $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

A major controversy in probability theory is over the types of statements to which probabilities can be assigned.

Frequentist Probability

The "Frequentists" restrict the assignment of probabilities to statements that describe the outcome of an experiment that can be repeated.

Let A denote the occurrence of head in one toss of a fair coin. Then, $P(A) = \lim_{n \to \infty} \frac{n(A)}{n}$, where n denotes the number of coin toss and $n(A)$ denotes the number of outcomes/trials favourable to the event A.

A problem or disadvantage with such a definition of probability is that it requires an infinite number of trials.

Subjective Probabilities

Within this school of thought, the belief is that probability theory is applicable to any situation in which there is uncertainty. This idea of probability is much more general. Outcomes of repeated experiments fall in this category, but so do statements about tomorrow's weather.

Calling the probabilities "subjective" does not imply that they may be assigned without regard to the axioms of probability. *De Finetti (1990)* provides a principle for assigning probabilities that does not rely on outcomes of repeated experiments but is consistent with probability axioms.

Lets try to understand the implications for statistical inference of adopting a subjective view of probability through a simple example.

Let $y = 1$ if a coin toss results in a head and 0 otherwise, and let $P(y = 1) = \theta$. So, $y \sim Ber(\theta)$. We are interested in learning about the parameter θ from an experiment in which the coin is tossed n times yielding the data $y = (y_1, y_2, \cdots, y_n)$

Frequentist: Probability theory can tell something about the distribution of data for a given θ because that data can be regarded as the outcome of a large number of repetitions of tossing a coin n times. The parameter θ is unknown number between 0 and 1.

Subjectivist: Here θ is an unknown quantity. Since there is uncertainty over its values, it can be regarded as a random variable and assigned a probability distribution. Before seeing the data, it is assigned a prior distribution $\pi(\theta), 0 \leq \theta \leq 1$. After accounting for data, it has the posterior distribution denoted by $\pi(\theta|y)$ calculated using the Bayes theorem.

1. Discrete: $\pi(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{p(y)}$ $\qquad$ $p(y) = \sum_\theta p(y|\theta)\pi(\theta)\, d\theta$
2. Continuous: $\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)}$ $\qquad$ $f(y) = \int f(y|\theta)\pi(\theta)\, d\theta$

The relationship between the posterior distribution, likelihood, and prior distribution is summarized by the equation,

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)} \tag{1}$$

or $\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$ in the unnormalized form. This follows from the Bayes' Theorem and forms the basis of Bayesian econometrics/statistics.

General notations to be remembered:

(1) $\pi(\theta)$: prior distribution on the parameters.

(2) $f(y|\theta)$: likelihood function. This is the joint pdf viewed as function of parameters.

(3) $\pi(\theta|y)$: posterior distribution.

(4) $f(y)$: marginal likelihood, obtained by integrating the numerator. Normalized the distribution.

Here, it is important to note that the likelihood function is not a *pdf* for θ. In particular, its integral over θ is not equal to 1, although its integral (or sum) over y is 1.

The unnormalized form $\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$ can be used as a method for updating information.

Example 1: Beta-Bernoulli Model

Suppose, we have n observations from the model $y_i \sim Ber(\theta)$ given by $y = (y_1, \cdots, y_n)'$. Then the likelihood can be written as,

$$P(y_1, \cdots, y_n; \theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}. \tag{2}$$

Let the prior distribution on θ be a Beta distribution (i.e., $\theta \sim Beta(\alpha, \beta)$) with the following *pdf*,

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{3}$$

Then the posterior distribution is obtained by combining the prior distribution with the likelihood using the Bayes Theorem as follows:

$$\begin{aligned}
\pi(\theta|y) &\propto p(y|\theta)\pi(\theta) \\
&\propto \theta^{\sum y_i}\,(1-\theta)^{n-\sum y_i}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1} \\
&\propto \theta^{(\alpha+\sum y_i)-1}\,(1-\theta)^{n-\sum y_i-1+\beta} \\
&\propto \theta^{\tilde{\alpha}-1}(1-\theta)^{\tilde{\beta}-1},
\end{aligned}$$

where $\tilde{\alpha} = \alpha + \sum y_i$ and $\tilde{\beta} = n + \beta - \sum y_i$. The last expression is recognized as the kernel of a Beta distribution and hence the posterior distribution is a Beta distribution. The posterior distribution summarizes all available information about θ, both from what was known before obtaining the current data and from the current data y.

Note that both the prior and posterior distribution are beta distributions. This is an example of conjugate prior. Beta distribution is a conjugate prior to Bernoulli likelihood.

Note that α can be interpreted as "the number of heads obtained in the 'experiment' on which the prior is based." Similarly, β represents the number of tails in the 'experiment' on which the prior is based.

We can summarize the posterior distribution by the posterior mean and posterior variance. Using the properties of the Beta distribution, the posterior mean is,

$$\begin{aligned}
E(\theta|y) &= \frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}} = \frac{\alpha+\sum y_i}{\alpha+\beta+n} \\
&= \Big(\frac{\alpha+\beta}{\alpha+\beta+n}\Big)\frac{\alpha}{\alpha+\beta} + \Big(\frac{n}{\alpha+\beta+n}\Big)\bar{y}.
\end{aligned}$$

The expression in the last line shows that the posterior mean is actually a weighted average of the prior means $\frac{\alpha}{\alpha+\beta}$ and maximum likelihood (ML) estimator $\bar{y}$.

The posterior variance is,

$$Var(\theta|y) = \frac{E(\theta|y)\big[1-E(\theta|y)\big]}{\alpha+\beta+n+1}.$$

Chapter 3: Posterior Distribution and Inference

Employing the Bayes Theorem', the posterior distribution is obtained by multiplying the likelihood with the prior distributions.

We now extend this process to include models with more than one parameters. We then discuss the revision of posterior distributions: (i) as more data becomes available, (ii) the role of the sample size, and (iii) the concept of identification.

Notations and Terminologies: We will use $\theta = (\theta_1, \theta_2, \cdots, \theta_d)$ to denote the vector of parameters. Given that θ is a vector, prior distribution will be joint prior distribution, likelihood becomes joint likelihood, and posterior distribution is referred to as joint posterior distribution.

From a joint distribution, we may derive marginal and conditional distributions according to the usual rules of probability. Suppose, we are primarily interested in the parameter θ_1.

$$\text{Marginal Posterior:} \quad \pi(\theta_1|y) = \int \pi(\theta_1, \cdots, \theta_d|y)\, d\theta_2, \cdots, d\theta_d$$

$$\text{Conditional Posterior:} \quad \pi(\theta_1|\theta_2, \cdots, \theta_d, y) = \frac{\pi(\theta_1, \cdots, \theta_d|y)}{\pi(\theta_2, \cdots, \theta_d|y)}$$

where the denominator on the RHS of the second equation is the marginal posterior of distribution $(\theta_2, \cdots, \theta_d)$ obtained by integrating out θ_1 from the joint distribution. The marginal distribution of a parameter is more useful, because marginal takes into account the uncertainty over the values of the remaining parameters, while the conditional distribution sets the remaining parameters at particular values.

Marginal distributions, while easy to write analytically, can be difficult to calculate as it involves multi-dimensional integration. The difficulty especially surfaces when the integral is not of a standard form.

Example 2: Dirichlet-Multinomial Model Let $y_i \sim MN(\theta_1, \cdots, \theta_d)$, where $\sum_{i=1}^{n} \theta_i =$

1. Let the experiment be repeated n times and outcome i arises y_i times. The likelihood function is as follows:

$$P(y_1, \cdots, y_n|\theta_1, \cdots, \theta_d) \quad = \quad \theta_1^{y_1} \cdots \theta_d^{y_d} \qquad \text{where } \sum y_i = n.$$

A simple example is the toss of a single die, for which $d = 6$. If the die is fair, $\theta_i = \frac{1}{6}$ for each possible outcome.

Prior distribution: To keep calculations simple, we use a conjugate prior distribution that generalizes the Beta distribution employed for the Bernoulli model. This is the Dirichlet distribution denoted $\theta \sim D(\alpha_1, \cdots, \alpha_d)$ with *pdf* as follows:

$$\pi(\theta_1, \cdots, \theta_d) = \frac{\Gamma(\sum_{i=1}^{d} \alpha_i)}{\prod_{i=1}^{d} \Gamma(\alpha_i)}\, \theta_1^{\alpha_1 - 1} \cdots \theta_d^{\alpha_d - 1} \qquad \alpha_i > 0,\ \sum \theta_i = 1$$

where $\alpha = (\alpha_1, \cdots, \alpha_d)$. The α_i are chosen to represent prior beliefs about the likely values of the θ_i's. As in the Bernoulli model, α_i can be interpreted as the number of times outcome i has appeared in previous experiments and $\sum \alpha_i$ represents the total number of trials on which prior is based.

One may set $\alpha_i = \alpha$ which implies a symmetrical treatment for each outcome. Setting $\sum \alpha_i$ to a small value is equivalent to weak prior information.

Using Bayes Theorem, the posterior distribution can be written as,

$$\begin{aligned} \pi(\theta|y) &\propto \theta_1^{\alpha_1-1}, \cdots, \theta_d^{\alpha_d-1}\theta_1^{y_1} \cdots \theta_d^{y_d} \\ &\propto \theta_1^{\alpha_1+y_1-1} \cdots \theta_d^{\alpha_d+y_d-1} \end{aligned}$$

This is recognized as the kernel of a Dirichlet distribution, so, $\theta|y \sim D(\alpha_1 + y_1, \cdots, \alpha_d + y_d)$ or more compactly $\theta|y \sim D(y + \alpha)$. The Dirichlet distribution is a conjugate prior for the multinomial model.

The marginal distribution for any of the θ_i, for example θ_1, can be derived and is,

$$\pi(\theta_1|y) \sim Beta\left(y_1 + \alpha_1, \sum_{i \neq 1}(y_i + \alpha_i)\right).$$

In the die throwing example, the probability of Spot 1 appearing when a single die is thrown is given by the Beta distribution,

$$\theta_1|y \sim Beta\left(y_1 + \alpha_1, \sum_{i=2}^{6}(y_i + \alpha_i)\right).$$

This is equivalent to considering the Spot 1 as one outcome and the remaining die faces as a second outcome, transforming the multinomial model into a binomial model.

Bayesian Updating

An interesting feature of Bayesian methods is Bayesian updating. The posterior distribution is obtained by combining the likelihood with the prior distribution using Bayes Theorem. The posterior distribution can be updated as new information becomes available, without actually starting from the beginning.

Let θ represent the parameter (scalar/vector) and let y_1 represent the fist set of data obtained in an experiment. As usual,

$$\pi(\theta|y_1) \propto f(y_1|\theta)\pi(\theta)$$

Now, suppose a new set of data y_2 is obtained and we want to compute the posterior distribution

using the complete data. Therefore,

$$\begin{aligned}\pi(\theta|y_1, y_2) &\propto f(y_1, y_2|\theta)\pi(\theta) \\ &\propto f(y_2|y_1, \theta)f(y_1|\theta)\pi(\theta) \\ &\propto f(y_2|y_1, \theta)\pi(\theta|y_1)\end{aligned}$$

If the datasets are independent, then $f(y_2|y_1, \theta) = f(y_2|\theta)$

Example: Consider the Bernoulli model with $Beta(\alpha, \beta)$ priors. Let the first experiment produce n_1 trials and let $S_1 = \sum y_{1i}$. Similarly, let the second experiment produce n_2 trials and let $S_2 = \sum y_{2i}$. Posterior distribution base on first experiment:

$$\begin{aligned}f(\theta|s_1) &\propto \theta^{\alpha-1}\,(1-\theta)^{\beta-1}\,\theta^{s_1}\,(1-\theta)^{n_1-s_1} \\ \theta|s_1 &\propto Beta(\alpha + s_1, \beta + (n_1 - s_1))\end{aligned}$$

Bayesian Updating: If we take the last expression as the prior for the second experiment, then,

$$\begin{aligned}f(\theta|s_1, s_2) &\propto \theta^{s_1+\alpha-1}\,(1-\theta)^{n_1+\beta-s_1-1}\,\theta^{s_2}\,(1-\theta)^{n_2-s_2} \\ \text{or,} \qquad \theta|s_1, s_2 &\sim Beta\Big(\alpha + (s_1 + s_2), \beta + (n_1 + n_2) - (s_1 + s_2)\Big).\end{aligned}$$

The latter distribution is implied by specifying a $B(\alpha_1, \beta_1)$ prior and obtaining $s_1 + s_2$ ones on $n_1 + n_2$ trials, where $\alpha_1 = \alpha + s_1 + s_2$ and $\beta_1 = \beta + (n_1 + n_2) - (s_1 + s_2)$.

To summarize, when data are generated sequentially, the Bayesian paradigm shows that the posterior distribution for the parameter based on new evidence is proportional to the likelihood based on the new data, and the prior distribution is actually the posterior distribution from the old data.

Large Samples

Bayesian inference holds true for any sample size. However, it is interesting to examine how posterior distribution behaves in large samples - particularly the relative contribution of the priors distributions and likelihood function in determining the posterior distribution. We explore this here.

Assuming the trials are independent, the likelihood and log likelihoods can be written as:

$$\begin{aligned}L(\theta, y) &= \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} L(\theta; y_i) \\ l(\theta; y) &= \ln L(\theta; y) = \sum l(\theta; y_i) = n\bar{l}(\theta; y)\end{aligned}$$

where $l(\theta; y_i)$ is the log-likelihood contribution of y_i and $\bar{l}(\theta; y) = \frac{1}{n}\sum l(\theta; y_i)$ is the mean log-likelihood contribution.

We re-express the posterior distribution as follows,

$$\begin{aligned}\pi(\theta|y) &\propto \pi(\theta)L(\theta;y)\\ &= \pi(\theta)\exp(n\bar{l}(\theta;y))\end{aligned}$$

From above, it is clear that as n increases, the exponential term grows and the relative contribution of data increase, while that of prior decreases.

The dominance of data can be taken to one step further. Let θ_0 denote the true value of θ, then it can be shown,

$$\lim_{n\to\infty} \bar{l}(\theta;y) \to \bar{l}(\theta_0;y)$$

i.e. posterior distribution collapses to a distribution with all its probability at θ_0. This is similar to the "consistency" property in the frequentist literature.

To derive the form of the posterior distribution for large n, we take a second order Taylor series approximation of $l(\theta;y)$ around $\hat{\theta}$, the MLE of θ:

$$\begin{aligned}l(\theta;y) &= l(\hat{\theta};y) - \frac{n}{2}(\theta-\hat{\theta})^2\big[-\bar{l}''(\hat{\theta};y)\big]\\ &= l(\hat{\theta};y) - \frac{n}{2\nu}(\theta-\hat{\theta})^2\end{aligned}$$

where $\bar{l}''(\hat{\theta};y) = \frac{1}{n}\sum_k l''(\hat{\theta};y_k)$ and $\nu = [-\bar{l}''(\hat{\theta};y)]^{-1}$.

The first derivative vanishes because $l(\theta;y)$ is maximised at $\theta=\hat{\theta}$. The posterior distribution, therefore, reduces to,

$$\pi(\theta|y) \propto \pi(\hat{\theta})\exp\left[-\frac{n}{2\nu}(\theta-\hat{\theta})^2\right]$$

which is a normal distribution with mean $\hat{\theta}$ and variance $\frac{v}{n}$, if $\pi(\hat{\theta}) \neq 0$. This is important. It implies that the prior distribution should not rule out values which are logically possible. Such values of θ may be strongly favoured by the likelihood function, but would have zero posterior probability if $\pi(\hat{\theta}) = 0$.

Multiparameter Case:

$$\begin{aligned}l(\theta;y) &= l(\hat{\theta};y) - \frac{n}{2}(\theta-\hat{\theta})'\big[-\bar{l}''(\hat{\theta};y)\big](\theta-\hat{\theta})\\ &= l(\hat{\theta};y) - \frac{n}{2}(\theta-\hat{\theta})'V^{-1}(\theta-\hat{\theta})\end{aligned}$$

where $\bar{l}''(\hat{\theta};y) = \frac{1}{n}\sum_k\left[\frac{\partial^2 l(\hat{\theta};y_k)}{\partial\theta_i\partial\theta_j}\right]$ is the mean of the matrix of second derivatives of the log likelihood evaluated at the MLE and $V = [-\bar{l}''(\hat{\theta};y)]^{-1}$. So, for large n,

$$\pi(\theta|y) = N\Big(\hat{\theta}, V/n\Big).$$

Summary: When n is large,

1. The prior distribution plays a relatively small role in determining the posterior.
2. The posterior distribution converges to a degenerate distribution at the true value of the parameter.
3. The posterior distribution is approximately normally distributed with mean $\hat{\theta}$.

Identification

Two models are said to be observationally equivalent, if $f(y|\theta) = f(y|\psi)$ i.e. the likelihoods are equivalent where θ and ψ are different parameters. In such a case, we cannot say or determine which model has generated the data.

The model or parameters of the model are not identified or unidentified when two or more models are observationally equivalent. The model is identified if no model is observationally equivalent to the model of interest.

A special case of non-identifiability arises when $f(y|\theta_1, \theta_2) = f(y|\theta_1)$. In such cases, the parameter θ_2 is not identified.

Example: In a linear regression that has categorical variables, we include indicator variables for $(k-1)$ categories. Including indicators for k categories makes the intercept perfectly collinear with the set of dummmies - this is a symptom of non identifiability of the constant and the coefficients of a complete set of dummies.

The identification described above is "identification through the data" but the Bayesian approach also utilizes the prior distribution. Consider the likelihood function $f(y|\theta_1, \theta_2) = f(y|\theta_1)$. There is no information about θ_2 given θ_1. What can we say about $\pi(\theta_2|y)$?

$$\begin{aligned}
\pi(\theta_2|y) &= \int \pi(\theta_1, \theta_2|y) d\theta_1 \\
&= \frac{1}{f(y)} \int f(y|\theta_1, \theta_2)\pi(\theta_1)\pi(\theta_2|\theta_1) d\theta_1 \\
&= \frac{1}{f(y)} \int f(y|\theta_1)\pi(\theta_1)\pi(\theta_2|\theta_1) d\theta_1 \\
&= \int \pi(\theta_1|y)\pi(\theta_2|\theta_1) d\theta_1 \qquad \text{since, } \pi(\theta_1|y) = \frac{f(y|\theta_1)\pi(\theta_1)}{f(y)}.
\end{aligned}$$

If $\pi(\theta_2|\theta_1) = \pi(\theta_2)$ then $\pi(\theta_2|y) = \pi(\theta_2)$, which implies a knowledge of y does not modify beliefs about θ_2. However, if $\pi(\theta_2|\theta_1) \neq \pi(\theta_2)$, the information about y modifies beliefs about θ_2 by modifying beliefs about θ_1.

Take home point: The researchers should know whether the parameters included in a model are identified through the data or through the prior distributions when presenting and interpreting posterior distributions.

Inference

We discuss how posterior distribution serves as the basis for Bayesian statistical inference.

Point Estimates: Consider a scalar parameter θ. The Bayesian estimator of θ is the value that minimizes the expected loss function.

- Absolute Loss function: $L_1(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$
- Quadratic Loss function: $L_2(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$
- Bilinear Loss function: $L_3(\hat{\theta}, \theta) = \begin{cases} a|\hat{\theta} - \theta|, & \text{for } \theta > \hat{\theta} \\ b|\hat{\theta} - \theta|, & \text{otherwise.} \end{cases}$

where $a, b > 0$. For these loss functions, loss is minimised if $\hat{\theta} = \theta$ and increases as $|\hat{\theta} - \theta|$ increases.

The Bayesian estimator $\hat{\theta}$ is obtained by minimizing the following expected loss function:

$$E\big[L(\hat{\theta}, \theta)\big] = \int L(\hat{\theta}, \theta)\pi(\theta|y)d\theta$$

where the expectation is taken over the posterior distribution of θ. Under the quadratic loss, we minimize:

$$E\big[L(\hat{\theta}, \theta)\big] = \int (\hat{\theta} - \theta)^2 \pi(\theta|y)d\theta.$$

Differentiating with respect to $\hat{\theta}$ and setting the FOC to 0, we have,

$$\begin{aligned} 2\int (\hat{\theta} - \theta)\pi(\theta|y)d\theta &= 0 \\ \text{or,} \quad \hat{\theta} = \int \theta\pi(\theta|y)d\theta &= E(\theta|y). \end{aligned}$$

Therefore, under the quadratic loss function, the optimal point estimator is the mean posterior distribution of θ.

Emphasis: Both the frequentist and Bayesian approaches to point estimates involves an expected value. However, there are important differences.

Frequentist: Expectation is taken over the distribution of an estimator given the unknown parameter θ.

Bayesian: Expectation is taken over the posterior distribution of θ.

Example 3: Consider the coin tossing experiment. $y_i \sim Ber(\theta)$ and let $\hat{\theta} = \bar{y}$.

Frequentist: To determine whether $\hat{\theta}$ is unbiased, we find the distribution of $\bar{y}$ and compute its expected value over the distribution of $\bar{y}$,

$$E(\bar{y}) = \int \bar{y} f(\bar{y}|\theta)d\bar{y}$$

This considers every possible value of $\bar{y}$, not just the data that is observed. In contrast, Bayesian calculation is based on the posterior distribution conditioned only on the observed data.

There is one more important difference between classical and Bayesian estimation method.

Classical: There is no general method to find candidates for estimators. You propose one or two estimators and then see whether they satisfy specific criteria.

Bayesian: The procedure is rather mechanical: given a loss function, the problem is to find an estimator that minimizes expected loss.

Interval Estimates

Interval estimates can be reported apart from the point estimates. $P(\theta_L \leq \theta \leq \theta_U) = 0.95$, which implies that the probability of $\theta \in [\theta_L, \theta_U]$ is 0.95. These intervals are called Bayesian probability intervals or credibility intervals.

The Bayesian credible intervals are different from that of confidence intervals which make use of unobserved data and do not involve the probability distribution of a parameter. For example, consider the model $x_i \sim N(\mu, 1)$ for $i = 1(1)n$. Then, the 95 percent confidence interval $\left(\bar{x} - \frac{1.96}{\sqrt{n}}, \bar{x} + \frac{1.96}{\sqrt{n}}\right)$ follows from the result that 95 percent of all possible sample means $\bar{x}$ lie in the interval $\left(\mu - \frac{1.96}{\sqrt{n}}, \mu + \frac{1.96}{\sqrt{n}}\right)$. This calculation involves sample means that are not observed. This is in sharp contrast to the Bayesian approach which only makes use of the observed data (and does not use data that are not observed).

Prediction

Prediction requires us to find the predictive distribution.

Example 1 Continued:

Consider the coin tossing example, $y_i \sim Ber(\theta)$. Observed data: $y = (y_1, \cdots, y_n)'$. We wish to predict outcome for the next toss, y_{n+1}.

$$\begin{aligned} P(y_{n+1}|y) &= \int f(y_{n+1} = 1, \theta|y) d\theta \\ &= \int P(y_{n+1} = 1|\theta, y)\pi(\theta|y) d\theta \\ &= \int P(y_{n+1} = 1|\theta)\pi(\theta|y) d\theta \end{aligned}$$

The last line drops dependence on y because $y_i's$ are independent given θ. So, you calculate $P(y_{n+1}|y)$ by averaging over different values of θ, drawn from the posterior distribution.

In the general case, the predictive distribution for a new value y_f

$$f(y_f|y) = \int f(y_f|\theta, y)\pi(\theta|y) d\theta$$

where conditioning on y has been retained because y_f depends on y in some time series models.

Example 1 Continued:

In the coin tossing experiment, we found the posterior distribution in the form a Beta

distribution,

$$\begin{aligned}
\pi(\theta|y) &= \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}\theta^{\alpha_1-1}(1-\theta)^{\beta_1-1}, \qquad \text{where} \\
\alpha_1 = \tilde{\alpha} &= \alpha + \sum y_i \\
\beta_1 = \tilde{\beta} &= \beta + n - \sum y_i
\end{aligned}$$

Since, $P(y_{n+1}=1|\theta)=\theta$, we have,

$$\begin{aligned}
P(y_{n+1}=1|y) &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+\sum y_i)\Gamma(\beta+n-\sum y_i)}\int \theta\theta^{\alpha+\sum y_i-1}(1-\theta)^{\beta+n-\sum y_i-1}d\theta \\
&= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+\sum y_i)\Gamma(\beta+n-\sum y_i)}\int \theta^{\alpha+\sum y_i}(1-\theta)^{\beta+n-\sum y_i-1}d\theta \\
&= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+\sum y_i)\Gamma(\beta+n-\sum y_i)} \quad \frac{\Gamma(\alpha+\sum y_i+1)\Gamma(\beta+n-\sum y_i)}{\Gamma(\alpha+\beta+n+1)} \\
&= \frac{\alpha+\sum y_i}{\alpha+\beta+n} = E(\theta|y),
\end{aligned}$$

which is the mean of the posterior distribution.

Model Comparison

The main objective of model comparison is to select the best fitting model amongst a set of competing models. For the sake of simplicity, let us suppose there are two models:

Model 1: $\theta_1,\ \pi_1(\theta_1),\ f_1(y|\theta_1) \quad P(M_1)=P_1.$
Model 2: $\theta_2,\ \pi_2(\theta_2),\ f_2(y|\theta_2) \quad P(M_2)=P_2.$

The Bayesian approach is to compute $P(M_i|y)$ which is interpreted as "the probability that Model i is the correct model given the data." To compute the term $P(M_i|y)$, we utilize the Bayes theorem.

$$\begin{aligned}
P(M_1|y) &= \frac{P(M_1)f_1(y|M_1)}{f(y)} \\
&= \frac{P_1\int f_1(y,\theta_1|M_1)d\theta_1}{f(y)} \\
&= \frac{P_1\int f_1(y|\theta_1,M_1)\pi_1(\theta_1|M_1)d\theta_1}{f(y)}
\end{aligned}$$

where $f(y) = P_1\int f_1(y|\theta_1,M_1)\pi_1(\theta_1|M_1)d\theta_1 + P_2\int f_2(y|\theta_2,M_2)\pi_2(\theta_2|M_2)d\theta_2$. Each term of $f(y)$ contains the integral of a likelihood function with respect to a prior distribution,

$$m_i(y) = \int f_i(y|\theta_i,M_i)\pi_i(\theta_i|M_i)d\theta_i.$$

The above is known as the marginal likelihood for model i and is interpreted as the expected value of the likelihood function with respect to the prior distribution.

For comparing the two models, we compute the posterior odds ratio,

$$\begin{aligned} R_{12} = \frac{P(M_1|y)}{P(M_2|y)} &= \left(\frac{P_1}{P_2}\right) \left(\frac{\int f_1(y|\theta_1, M_1)\ \pi_1(\theta_1|M_1)\ d\theta_1}{\int f_2(y|\theta_2, M_2)\ \pi_2(\theta_2|M_2)\ d\theta_2}\right) \\ &= \left(\frac{P_1}{P_2}\right) \left(\frac{m_1(y)}{m_2(y)}\right) \\ &= \text{(Prior odds ratio)} \times \text{(Bayes factor)} \end{aligned}$$

When $R_{12} > 1$, M_1 is more supported by data and prior information compared to M_2. When $R_{12} \equiv 1$, both models are equally supported. *Drawback*: Only two models can be compared at a time.

When we have more than one model and prediction is required, we can employ 'model averaging'. Prediction may be done by a weighted average of the predictions of each model.

$$\begin{aligned} f(y_f|y) &= \sum_{i=1}^{m} P(M_i|y)\ f_i(y_f|y, M_i) \\ &= \sum_{i=1}^{m} P(M_i|y) \int f(y_f|\theta_i, y, M_i)\ \pi(\theta_i|y, M_i) d\theta_i \end{aligned}$$

Recall that $P(M_i|y)$ is interpreted as that M_i is the correct model given the data.

Unlike frequentist models, Bayesian model comparison can easily deal with non-nested hypothesis, including models that specify different representations of the response variable. To see the latter, suppose under M_1, the likelihood function is $f_1(y|\theta_1)$ and under M_2 it is $f_2(z|\theta_2)$, where $z = g(y)$ and $g(y)$ is monotone. Then, the posterior odds ratio is $\frac{P(M_1|y)}{P(M_2|y)}$ or $\frac{P(M_1|z)}{P(M_2|z)}$, and they are equivalent. This is because, by the usual transformation of variables rule,

$$f(z_i|\theta) = f(y_i|\theta)\left|\frac{dy_i}{dz_i}\right|$$

and therefore in the Bayes factor, the Jacobian cancels out. The result generalizes to multivariate y and z.

Effect of sample size on Bayes factor: A second order Taylor series approximation of the log-likelihood yields,

$$l(\theta; y) = l(\hat{\theta}; y) - \frac{n}{2}(\theta - \hat{\theta})' V^{-1} (\theta - \hat{\theta}), \tag{4}$$

where centering is around $\hat{\theta}$. We know the marginal likelihood has the expression,

$$m_i(y) = \int f_i(y|\theta_i, M_i)\pi_i(\theta_i|M_i)d\theta_i \tag{5}$$

Exponentiating (4) and substituting into (5), we have

$$\begin{aligned} m_i(y) &\simeq L_i(\hat{\theta}_i|y) \int \exp\left[-\frac{n}{2}(\theta_i - \hat{\theta}_i)' V^{-1} (\theta_i - \hat{\theta}_i) \right] \pi_i(\theta_i) d\theta_i \\ &\simeq L_i(\hat{\theta}_i|y) \pi_i(\hat{\theta}_i) \int \exp\left[-\frac{n}{2}(\theta_i - \hat{\theta}_i)' V^{-1} (\theta_i - \hat{\theta}_i) \right] d\theta_i \end{aligned}$$

where $\pi(\theta_i)$ is approximated by $\pi(\hat{\theta}_i)$ because the exponential term dominates the integral in the region around $\hat{\theta}_i$. The integration yields,

$$\begin{aligned} m_i(y) &\simeq L_i(\hat{\theta}_i|y)\ \pi_i(\hat{\theta}_i)(2\pi)^{\frac{d_i}{2}} \left| n^{-1} V_i \right|^{1/2} \\ &= L_i(\hat{\theta}_i|y)\ \pi_i(\hat{\theta}_i)(2\pi)^{\frac{d_i}{2}} n^{-\frac{d_i}{2}} |V_i|^{1/2}, \end{aligned}$$

where d_i is the dimension of θ_i, i.e. the number of parameters in model M_i

We can now approximate the logarithm of Bayes factor for comparing Model 1 and Model 2 as follows:

$$\begin{aligned} \ln(B_{12}) &\simeq \left[\ln\left(\frac{L_1(\hat{\theta}_1|y)}{L_2(\hat{\theta}_2|y)} \right) - (d_1 - d_2) \ln(n) \right] \\ &\quad + \left[\ln\left(\frac{\pi_1(\hat{\theta}_1)}{\pi_2(\hat{\theta}_2)} \right) + \frac{1}{2} \ln\left(\frac{|V_1|}{|V_2|} \right) + \frac{(d_1 - d_2)}{2} \ln(2\pi) \right]. \end{aligned}$$

The term $\log\left(\frac{L_1(\hat{\theta}_1|y)}{L_2(\hat{\theta}_2|y)} \right)$ will become large if M_1 is the true model and small if M_2 is true.

The expression $-(d_1 - d_2)\ln(n)$ shows that Bayes factor penalizes models with large number of parameters. Penalty is $\log(n)$ times the difference in number of parameters.

The second term in square bracket does not depend on n, its effect becomes small for large n. E.g.: Coin tossing experiment. See the book