Econometrics

Introduction to Bayesian Econometrics

November 2023 – January 2024

Why Bayesian econometrics?

- What does an econometrician do? i) Estimate parameters in a model (e.g. regression coefficients), ii) Compare different models (e.g. hypothesis testing), iii) Prediction
- Bayesian econometrics do all these based on a few simple rules of probability.

Bayesian Statistics

Key idea of Bayesian approach: The only satisfactory representation of uncertainty is through probability theory.

Bayesian Statistics

Key idea of Bayesian approach: The only satisfactory representation of uncertainty is through probability theory.

- **The Bayesian receipt**: Whatever is unknown and you want to estimate call it θ, whatever is known call it y. Then use probability theory to calculate $p(\theta|y)$.

Bayesian Statistics

Key idea of Bayesian approach: The only satisfactory representation of uncertainty is through probability theory.

- **The Bayesian receipt**: Whatever is unknown and you want to estimate call it θ, whatever is known call it y. Then use probability theory to calculate $p(\theta|y)$.
- Main **difference with classical statistics** (econometrics): θ is a random quantity/variable and not just a number as in the classical approach.

Bayesian Statistics

Key idea of Bayesian approach: The only satisfactory representation of uncertainty is through probability theory.

- **The Bayesian receipt**: Whatever is unknown and you want to estimate call it θ, whatever is known call it y. Then use probability theory to calculate $p(\theta|y)$.
- Main **difference with classical statistics** (econometrics): θ is a random quantity/variable and not just a number as in the classical approach.
- Bayesian estimation relies on $f(\theta|y)$ the distribution of θ given the observed data, whereas in the classical approach we rely on $f(y|\theta)$.

Bayesian Statistics

Key idea of Bayesian approach: The only satisfactory representation of uncertainty is through probability theory.

- **The Bayesian receipt**: Whatever is unknown and you want to estimate call it θ, whatever is known call it y. Then use probability theory to calculate $p(\theta|y)$.
- Main **difference with classical statistics** (econometrics): θ is a random quantity/variable and not just a number as in the classical approach.
- Bayesian estimation relies on $f(\theta|y)$ the distribution of θ given the observed data, whereas in the classical approach we rely on $f(y|\theta)$.
- Before we compute $f(\theta|y)$ we need to define $f(\theta)$ which is called the **prior distribution**.

Prior distribution

The prior distribution is the **core** of Bayesian statistics and is considered as the main advantage of those they prefer Bayesian estimation or the main disadvantage for the others.

- When we wish to estimate θ almost always we have some **knowledge or belief** for its possible values.

Prior distribution

The prior distribution is the **core** of Bayesian statistics and is considered as the main advantage of those they prefer Bayesian estimation or the main disadvantage for the others.

- When we wish to estimate θ almost always we have some **knowledge or belief** for its possible values.
- Assume for example one looks outside from the window and sees a wooden object with green leaves.

Prior distribution

The prior distribution is the **core** of Bayesian statistics and is considered as the main advantage of those they prefer Bayesian estimation or the main disadvantage for the others.

- When we wish to estimate θ almost always we have some **knowledge or belief** for its possible values.
- Assume for example one looks outside from the window and sees a wooden object with green leaves.
- There are two possible assumptions: the object is a tree or the object is a postman.

Prior distribution

The prior distribution is the **core** of Bayesian statistics and is considered as the main advantage of those they prefer Bayesian estimation or the main disadvantage for the others.

- When we wish to estimate θ almost always we have some **knowledge or belief** for its possible values.
- Assume for example one looks outside from the window and sees a wooden object with green leaves.
- There are two possible assumptions: the object is a tree or the object is a postman.
- We all think that it is a tree but let's see how this is translated in terms of probabilities: Let A the event that we see the wooden object, B_1 we consider as a tree and B_2 we consider as a post man.

Prior distribution

The prior distribution is the **core** of Bayesian statistics and is considered as the main advantage of those they prefer Bayesian estimation or the main disadvantage for the others.

- When we wish to estimate θ almost always we have some **knowledge or belief** for its possible values.
- Assume for example one looks outside from the window and sees a wooden object with green leaves.
- There are two possible assumptions: the object is a tree or the object is a postman.
- We all think that it is a tree but let's see how this is translated in terms of probabilities: Let A the event that we see the wooden object, B_1 we consider as a tree and B_2 we consider as a post man.
- We choose B_1 because intrinsically we calculate $f(A|B_1) > f(A|B_2)$. We need thus to include these intrinsic calculations in our estimation procedure.

Prior distribution

More intuition: In the following examples we are interested in estimating the probability of success.

1. We ask 10 times a woman from England to guess if there is milk in her tea and she gives 10 correct answers.
2. An experiences musician claims that he can classify a melody if is from Mozart or Vivaldi and he gives 10 correct answers.
3. A drunk man claims that he can guess between toss or coin and gives 10 correct answers.

In all the three cases the data suggest to estimate $\hat{p} = 1$ but do we "trust" the data in all the three cases?

The Bayes theorem

The main ingredient of Bayesian estimation is the Bayes theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Or more generally:

$$P(C_i|B) = \frac{P(B|C_i)P(C_i)}{\sum_{j=1}^{J} P(B|C_j)P(C_j)},$$

where $C_1, C_2, \ldots, C_J$ events that form a partition of a sample space Ω.

The Bayes theorem: Example

You are a financial analyst at an investment bank knowing that

- 60% of the publicly-traded companies increased their share price by more than 5% in the last 3 years replaced their CEO.
- For companies that didn't replace their CEO the proportion is 35%.
- Knowing that the probability that the stock prices grow by more than 5% is 4%, find the probability that the shares of a company that fires its CEO will increase by more than 5%.

$$P(A|B) = \frac{0.60 \times 0.04}{0.60 \times 0.04 + 0.35 \times (1 - 0.04)} = 0.067 \text{ or } 6.67\%$$

Figure: Probability that the shares of a company that replaces its CEO will grow by more than 5%.

Advanced Bayesian estimation

Basic steps to estimate the unknown θ based on data y:

1. Choose a likelihood model $f(y|\theta)$
2. Choose a prior distribution
3. From Bayes theorem find the posterior distribution $f(\theta|y)$
4. Make statistical inference. For example
   - Set $\hat{\theta}$ to be the mean of $f(\theta|y)$.
   - Set the 2.5% and 97.5% to form a **credible** (analogous to confidence) interval of $\alpha = 5\%$.

4.*

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

θ can be either continuous or discrete and $f(\theta)$ is pdf or pmf respectively.

Bayesian estimation: The denominator

The denominator of Bayes theorem is an integral wrt θ and thus for a given dataset y it does not depend on θ. Therefore, the Bayes theorem is also useful in the for

$$f(\theta|y) \propto f(\theta)f(y|\theta),$$

which are the only quantities in the posterior that depend on θ.

Choosing the prior distribution

Remark: $f(\theta)$ doesn't depend on data.

- Prior information is controversial aspect since it sounds unscientific.
- Bayesian answers (to be elaborated on later):
- i) Often we do have prior information and, if so, we should include it (more information is good)
- ii) Can work with "noninformative" priors
- iii) Can use hierarchical priors which treat prior hyperparameters as parameters and estimates them
- iv) Training sample priors
- v) Bayesian estimators often have better frequentist properties than frequentist estimators (e.g. results due to Stein show MLE is inadmissible – but Bayes estimators are admissible)
- vi) Prior sensitivity analysis

Bayesian predictions

- Prediction based on the *predictive density* $p(y^*|y)$
- Since a marginal density can be obtained from a joint density through integration:

$$p(y^*|y) = \int p(y^*, \theta|y) d\theta.$$

- Term inside integral can be rewritten as:

$$p(y^*|y) = \int p(y^*|y, \theta) p(\theta|y) d\theta.$$

- Prediction involves the posterior and $p(y^*|y, \theta)$ (more description provided later)

Bayesian Model Comparison

- Models denoted by M_i for $i = 1, .., m$. M_i depends on parameters θ^i.
- *Posterior model probability* is $p(M_i|y)$.
- Using Bayes rule with $B = M_i$ and $A = y$ we obtain:

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

- $p(M_i)$ is referred to as the *prior model probability*.
- $p(y|M_i)$ is called the *marginal likelihood*.

Bayesian Model Comparison

- How is marginal likelihood calculated?
- Posterior can be written as:

$$p(\theta^i|y, M_i) = \frac{p(y|\theta^i, M_i)p(\theta^i|M_i)}{p(y|M_i)}$$

- Integrate both sides with respect to θ^i, use fact that $\int p(\theta^i|y, M_i)d\theta^i = 1$ and rearrange:

$$p(y|M_i) = \int p(y|\theta^i, M_i)p(\theta^i|M_i)d\theta^i.$$

- Note: marginal likelihood depends only on the prior and likelihood.

Bayesian Model Comparison

- *Posterior odds ratio* compares two models:

$$PO_{ij} = \frac{p(M_i|y)}{p(M_j|y)} = \frac{p(y|M_i)p(M_i)}{p(y|M_j)p(M_j)}.$$

- Note: $p(y)$ is common to both models, no need to calculate.

Bayesian Model Comparison

- Can use fact that $p(M_1|y) + p(M_2|y) + ... + p(M_m|y) = 1$ and PO_{ij} to calculate the posterior model probabilities.
- E.g. suppose $m = 2$ models and you know:

$$p(M_1|y) + p(M_2|y) = 1$$

$$PO_{12} = \frac{p(M_1|y)}{p(M_2|y)}$$

- imply

$$p(M_1|y) = \frac{PO_{12}}{1 + PO_{12}}$$

$$p(M_2|y) = 1 - p(M_1|y).$$

- The *Bayes Factor* is:

$$BF_{ij} = \frac{p(y|M_i)}{p(y|M_j)}.$$

Advanced Bayesian Estimation: Example

- Background:
- Experiment repeated T times
- Each time the outcome can be "success" or "failure"
- y_t for $t = 1, .., T$ are random variables for each repetition of experiment
- Realization of y_t can be 1 or 0
- Probability of success is θ (hence probability of failure is $1 - \theta$)
- The goal is to estimate θ

Example: The likelihood model

- Notation for things above is: $y_t \in \{0, 1\}$, $0 \leq \theta \leq 1$ and

$$p(y_t|\theta) = \begin{cases} \theta & \text{if } y_t = 1 \\ 1 - \theta & \text{if } y_t = 0. \end{cases}$$

- Let m be the number of successes in T repetitions of experiment
- Likelihood function is:

$$\begin{aligned} p(y|\theta) &= \prod_{t=1}^{T} p(y_t|\theta) \\ &= \theta^m (1 - \theta)^{T-m} \end{aligned}$$

Example: The prior

- View this likelihood in terms of θ: proportional to p.d.f. of a Beta distribution
- See definition in textbook Appendix B or Wikipedia
- Most common distribution for random variables bounded to lie in the interval $[0, 1]$
- Commonly used for parameters which are probabilities (like θ)
- Bayesians need prior
- Let us also Beta distribution for prior
- Prior beliefs concerning θ are represented by

$$p(\theta) \propto \theta^{\underline{\alpha}-1}(1-\theta)^{\underline{\delta}-1}$$

Example: Prior Elicitation

- The researcher chooses prior hyperparameters $\underline{\alpha} > 0$ and $\underline{\delta} > 0$ to reflect beliefs
- Called prior elicitation
- Properties of Beta distribution imply prior mean is

$$E(\theta) = \frac{\underline{\alpha}}{\underline{\alpha} + \underline{\delta}}$$

- Suppose you believe, a priori, that success and failure are equally likely
- $E(\theta) = \frac{1}{2}$ obtained by setting $\underline{\alpha} = \underline{\delta}$
- If I look on Wikipedia I see $\underline{\alpha} = \underline{\delta} = 2$ has mean at $E(\theta) = \frac{1}{2}$ but spreads probability widely over interval $[0, 1]$
- So I might be "relatively noninformative" and choose this for my prior

Example: Prior Elicitation - Non-Informative

- Or I might set $\underline{\alpha} = \underline{\delta} = 1$ and be completely noninformative
- Note: $\underline{\alpha} = \underline{\delta} = 1$ implies $p(\theta) \propto 1$
- Uniform distribution over interval $[0, 1]$
- Every value for θ receives same probability (equally likely) = noninformative prior

Example: The posterior

- Posterior same Beta form as prior (terminology $=$ conjugate)
- Posterior has arguments $\overline{\alpha}$ and $\overline{\delta}$ instead of $\underline{\alpha}$ and $\underline{\delta}$
- Arguments have been updated:
- Begin with prior belief ($\underline{\alpha}$ or $\underline{\delta}$) update with data information (m and $T - m$)
- Posterior combines prior and data information
- "Bayesian learning" $=$ learn about θ by combining prior and data information

Example: The posterior

- To get posterior multiply prior times likelihood

$$\begin{aligned} p(\theta|y) &\propto \theta^{\underline{\alpha}-1}(1-\theta)^{\underline{\delta}-1}\theta^{m}(1-\theta)^{T-m} \\ &= \theta^{\overline{\alpha}-1}(1-\theta)^{\overline{\delta}-1} \end{aligned}$$

- where

$$\begin{aligned} \overline{\alpha} &= \underline{\alpha}+m \\ \overline{\delta} &= \underline{\delta}+T-m \end{aligned}$$

Bayesian Computation

- How do you present results from a Bayesian empirical analysis?
- $p(\theta|y)$ is a p.d.f. Especially if θ is a vector of many parameters cannot present a graph of it.
- Want features analogous to frequentist point estimates and confidence intervals.
- A common point estimate is the mean of the posterior density (or *posterior mean*).
- Let θ be a vector with k elements, $\theta = (\theta_1, .., \theta_k)'$. The posterior mean of any element of θ is:

$$E(\theta_i|y) = \int \theta_i p(\theta|y) d\theta.$$

Bayesian Computation

- Let $g()$ be a function, then the *expected value* of $g(X)$, denoted $E[g(X)]$, is defined by:

$$E[g(X)] = \sum_{i=1}^{N} g(x_i)\, p(x_i)$$

- if X is discrete random variable with sample space $\{x_1, x_2, x_3, .., x_N\}$
- $$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, p(x)\, dx$$
- if X is a continuous random variable (provided $E[g(X)] < \infty$).

Bayesian Computation

- Common measure of dispersion is the *posterior standard deviation* (square root of *posterior variance*)
- Posterior variance:

$$var(\theta_i|y) = E(\theta_i^2|y) - \{E(\theta_i|y)\}^2,$$

- This requires calculating another expected value:

$$E(\theta_i^2|y) = \int \theta_i^2 p(\theta|y) d\theta.$$

- Many other possible features of interest. E.g. what is probability that a coefficient is positive?

$$p(\theta_i \geq 0|y) = \int_0^\infty p(\theta_i|y) d\theta_i$$

Bayesian Computation

- All of these posterior features have the form:

$$E\left[g\left(\theta\right)|y\right] = \int g(\theta)p(\theta|y)d\theta,$$

- where $g(\theta)$ is a *function of interest*.
- All these features have integrals in them. Marginal likelihood and predictive density also involved integrals.
- Apart from a few simple cases, it is not possible to evaluate these integrals analytically, and we must turn to the computer.

Bayesian Computation

- The integrals involved in Bayesian analysis are usually evaluated using simulation methods.
- Will use several methods later on. Here we provide some intuition.
- Frequentist asymptotic theory uses Laws of Large Numbers (LLN) and a Central Limit Theorems (CLT).
- A typical LLN: "consider a random sample, $Y_1, ..Y_N$, as N goes to infinity, the average converges to its expectation" (e.g. $\overline{Y} \to \mu$)
- Bayesians use LLN: "consider a random sample from the posterior, $\theta^{(1)}, ..\theta^{(S)}$, as S goes to infinity, the average of these converges to $E\left[\theta|y\right]$"
- Note: Bayesians use asymptotic theory, but asymptotic in S (under control of researcher) not N